# Methodology
# Lesson 3: Preparing the data & experiments

---

Carlos Ramisch
first.last@lis-lab.fr

M2 IAAA - based on the course *Zen Research*
By Carlos Ramisch and Manon Scholivet

# Why do we need experiments?

- A research question and its sub-questions
  - $\longrightarrow$ Precise, concise, feasible, interesting
- Hypotheses related to each sub-question
- They are anchored in the litterature and justified

# Why do we need experiments?

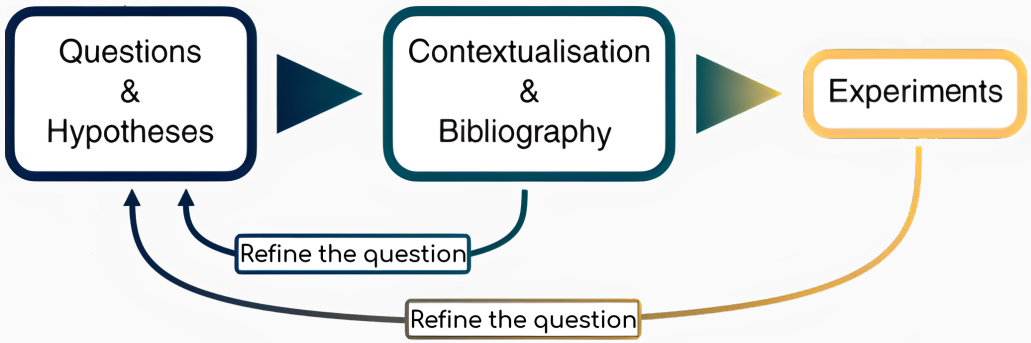
- A research question and its sub-questions
  - $\longrightarrow$ Precise, concise, feasible, interesting
- Hypotheses related to each sub-question
- They are anchored in the litterature and justified

**Experiment goals**

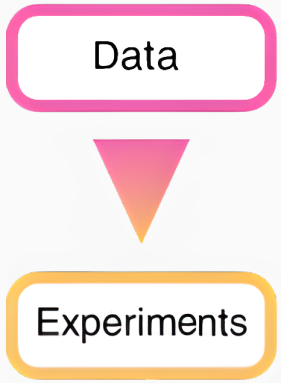
1. To build further evidence to support hypotheses
   - $\longrightarrow$ Eventually end up accepting or rejecting them
2. Lead to new interesting research questions

# Remember: ideal vs. reality

## Ideally:




## In practice:



# Designing an experiment

1. Identify the target hypothesis

   → Prioritise hypotheses according to impact and constraints

# Designing an experiment

1. Identify the target hypothesis

   → Prioritise hypotheses according to impact and constraints

2. Identify the needs of the experiment

   → Data, datasets, evaluation metrics

# Designing an experiment

1. Identify the target hypothesis
   → Prioritise hypotheses according to impact and constraints
2. Identify the needs of the experiment
   → Data, datasets, evaluation metrics
3. Instantiate under-specified aspects of the question/hypotheses
   → The devil is in the details

# Designing an experiment

1. Identify the target hypothesis
   → Prioritise hypotheses according to impact and constraints
2. Identify the needs of the experiment
   → Data, datasets, evaluation metrics
3. Instantiate under-specified aspects of the question/hypotheses
   → The devil is in the details
4. If the result is X, I will be able to conclude Y
   → Reformulate hypotheses in terms of experiment outcomes

# Refining the hypothesis: example

**Hypothesis**

It is possible to learn a model for language $L$ (with no annotations available) from a set of languages $L'$ (with available annotations)

# Refining the hypothesis: example

**Hypothesis**

It is possible to learn a model for language $L$ (with no annotations available) from a set of languages $L'$ (with available annotations)

- A model for which task? Question answering? Parsing?
  - $\rightarrow$ What can the model learn from annotated data?
- What exact set of languages?
- What configurations will be tested?
  - $\rightarrow$ $L'$ contains 1 language, 5 languages...
  - $\rightarrow$ $L$ is similar to a language in $L'$ or not?
- How to assess if a model is good? Evaluation metrics?

# Wooclap

Wooclap time!

# Experimental protocol

- Step-by-step description of the experiment
- "Algorithm" of the experiment

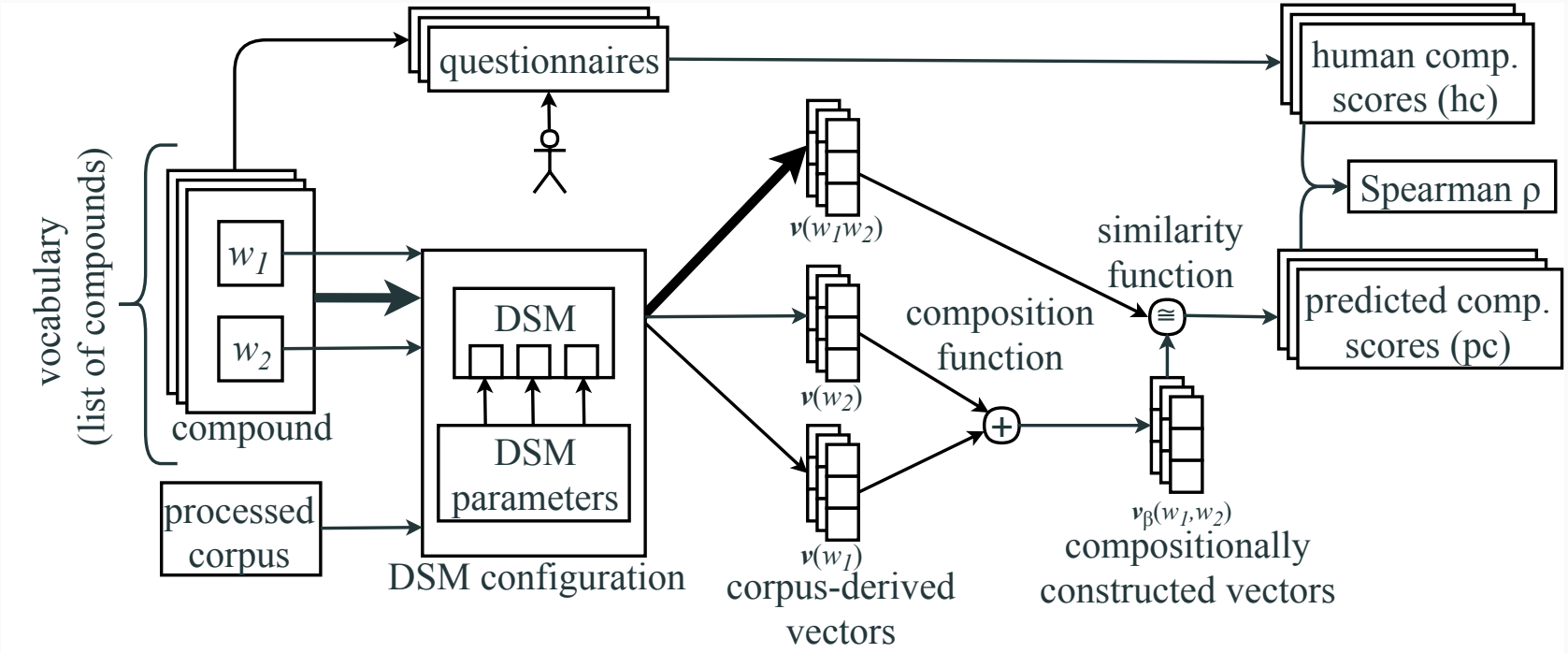
  $\longrightarrow$ Writing the recipe before start cooking

# Experimental protocol document

How formal is the protocol description?

- Depends on the discipline
  - → E.g. medicine: protocol published before doing the study
  - → Ethics committee (human/animal subjects, biases, . . . )
- A good protocol description can speed up paper writing
  - → Can be a schema, kanban, Gantt diagram,. . .
- In any case, to be defined before launching experiments

# Experimental protocol diagram: example




Source: `https://aclanthology.org/J19-1001/`

# Making choices

- Beware of the combinatorial explosion
  - → # datasets × # configs × # models × # metrics × ...
- Choices must be justified
  - → An arbitrary justification is better than none
  - → E.g. *the parameter was chosen by trial and error*
- Favour more promising aspects
  - → Similar metrics → choose one vs. varied datasets → test all
  - → Small pilot experiments ⟹ trends ⟹ choices

# Outline

Dataset creation (annotation)

Data quality metrics (agreement)

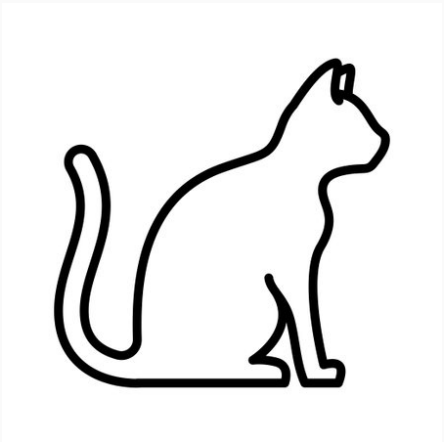
Experiments management

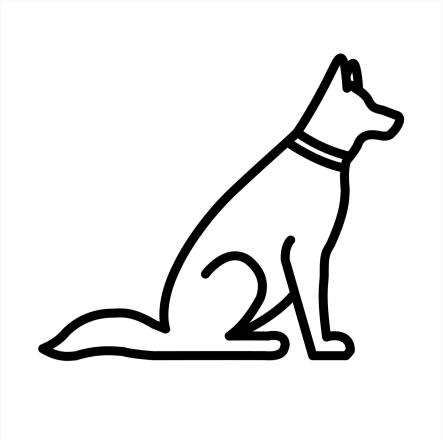
Data management

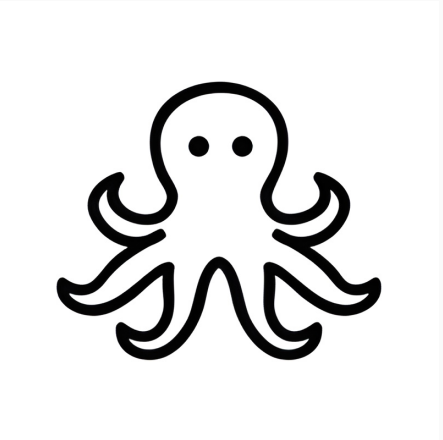
# General context: supervised machine learning

- Supervised methods require
  - Input $x$ associated with expected output $y$

Input







Reference Cat Dog Octopus

- gold = reference = label = ground truth

# Where does data come from?

- Machine learning courses:

```
from sklearn.datasets import load_digits
digits = load_digits()
print(digits.target[:20])    # magic !
[0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9]
```

# Where does data come from?

- Machine learning courses:

```
from sklearn.datasets import load_digits
digits = load_digits()
print(digits.target[:20])    # magic !
[0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9]
```

- Real life: Here's some data ($x$), apply some ML on it!

  $\rightarrow$ How to obtain gold/reference labels $y$ to learn/evaluate models?

# Where does data come from?

- Machine learning courses:

```
from sklearn.datasets import load_digits
digits = load_digits()
print(digits.target[:20])    # magic !
[0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9]
```

- Real life: Here's some data ($x$), apply some ML on it!

  $\rightarrow$ How to obtain gold/reference labels $y$ to learn/evaluate models?

- **Lucky day:** perfect dataset already available!
- **Other days:** create dataset for my needs

  $\rightarrow$ Or maybe: dataset exists but needs extensions/improvements

# Dataset creation recipe

Step 1: Select or collect material to annotate

Step 2: Write annotation guidelines

Step 3: Develop or adapt an annotation interface

Step 4: Recruit and train annotators

Step 5: Evaluate quality of annotations

Step 6: Aggregate annotations

Step 7: Format, document, release [optional]

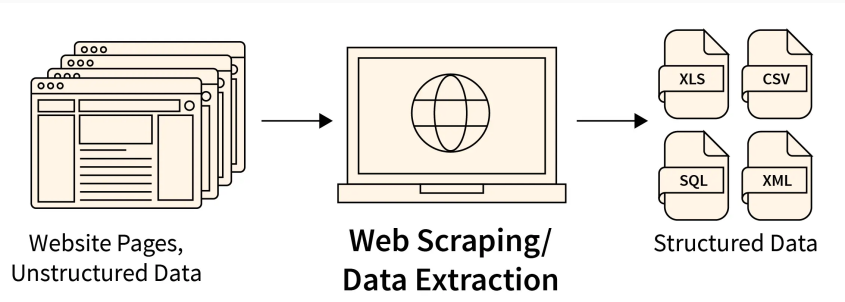
$\longrightarrow$ Use dataset for experiments!

# Step 1: Data selection for annotation

- Similarity with target application data
- Trade-off between realistic vs. artificial
  - → E.g. newspaper vs. tweets
  - → Climate crisis means quarter of ski resorts face scarce snow
  - → sooo sick of the snow UGHH!!! ò.ó
- Raw data is noisy $\Longrightarrow$ harder to annotate/exploit
  - → E.g. for text: dialects, typos, code switching, slang...

# Example: Text crawling / scraping[1]

- Obtain data (HTML) from the web
  - → Off-the-shelf tools, e.g. BootCat
  - → Pre-downloaded web dumps: CommonCrawl, Wikimedia
  - → In-house scripts: parallelisation, `robots.txt`, priority queue ...
- Pre-processing and cleaning
  - → Language identification, e.g. LangID / deduplication, e.g. Onion
  - → Boilerplate removal, e.g. BeautifulSoup, jusText
  - → Content filtering, e.g. regexps / segmentation, e.g. Spacy, NLTK




[1] Extracting data (e.g. text) from websites using a software (e.g. crawler, bot).

# Indirect annotation

Clever data selection - “free” annotations

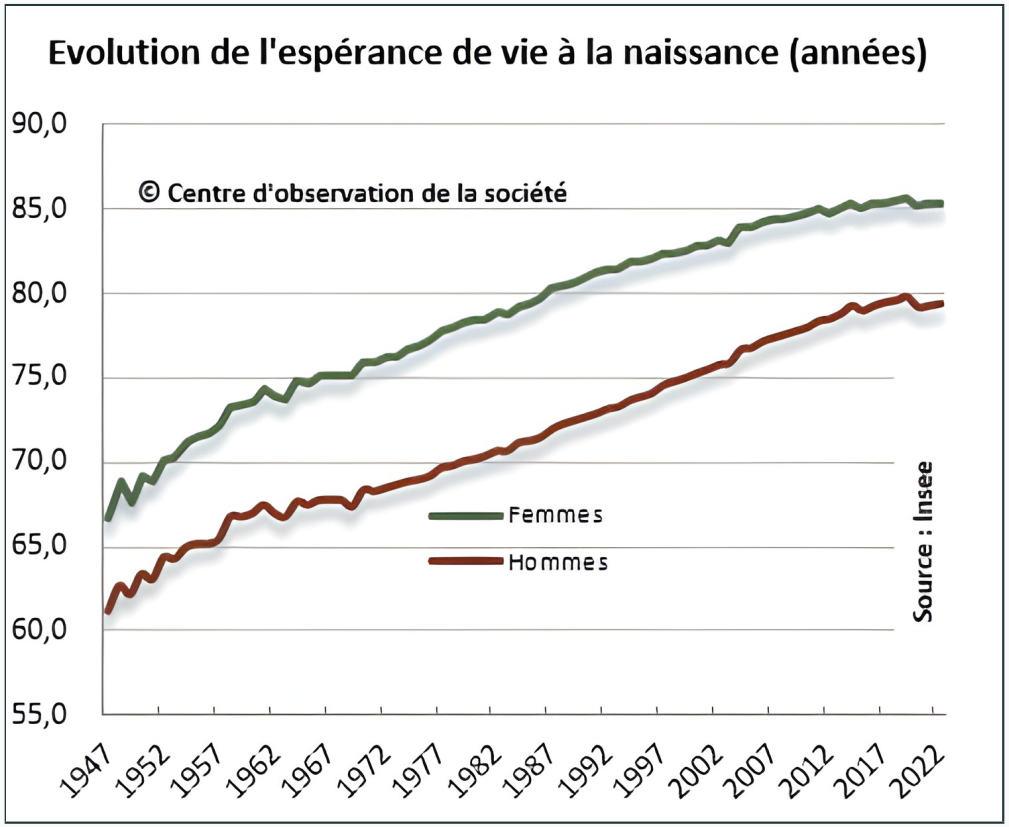
- Open Subtitles: text + translation provided by cinema fans
- Amazon reviews: text + 5-star rating provided by customers
- Flickr30k: image + captions provided by FLickr users
- . . .


# Warning!

# Selection bias

Selecting data completely at random is actually quite hard

- E.g. study the average lifetime of human beings
- Oops! 90% of the dataset consists of men...

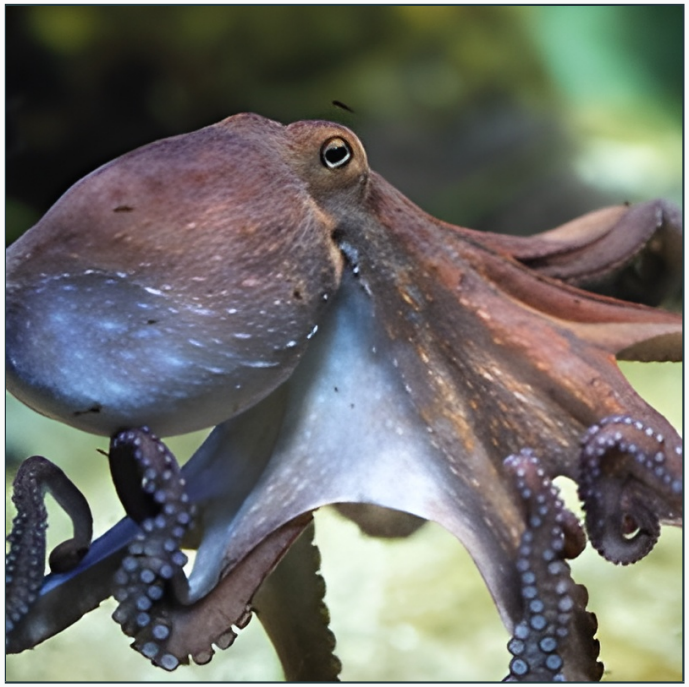


# Wooclap

Wooclap time!

# Annotation example: image classification

- Is this an octopus? Yes/No



# Annotation example: image classification

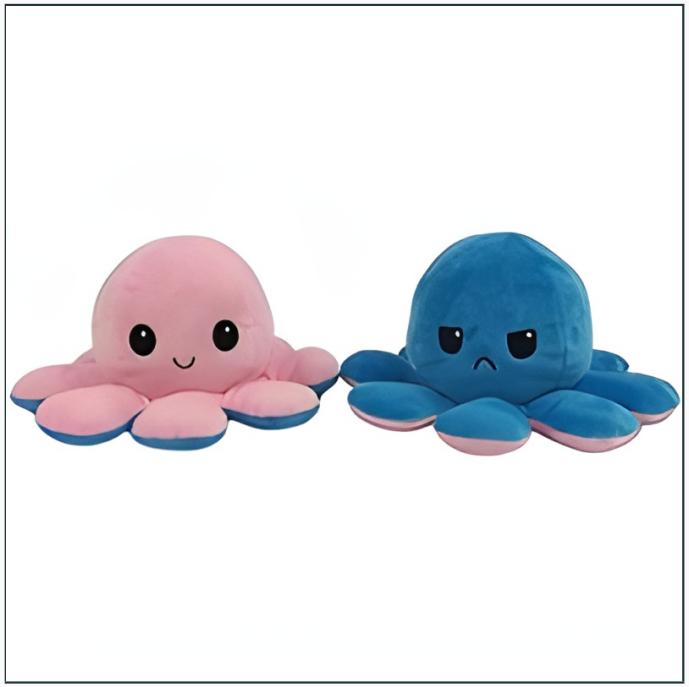
- Is this an octopus? Yes/No

# Annotation example: image classification

- Is this an octopus? Yes/No

# Annotation example: image classification

- Is this an octopus? Yes/No

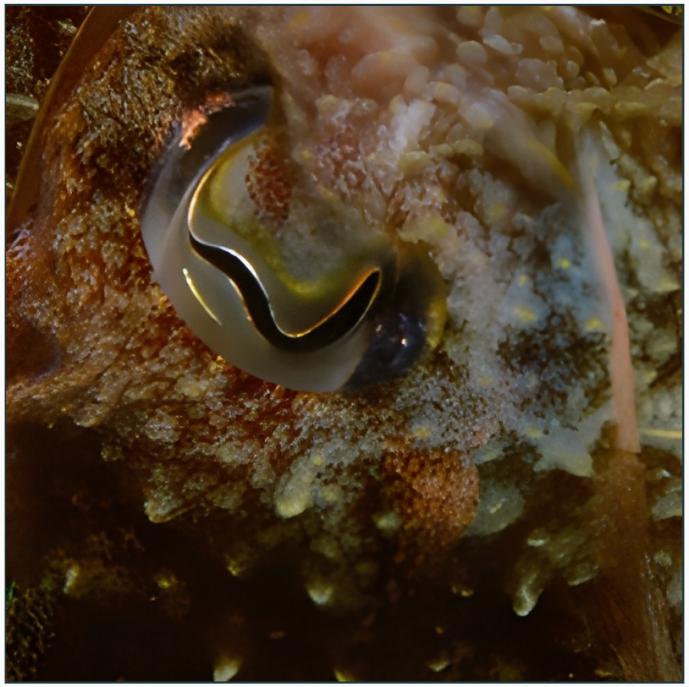


# Annotation example: image classification

- Is this an octopus? Yes/No

# Annotation example: image classification

- Is this an octopus? Yes/No



# Step 2: Annotation guidelines

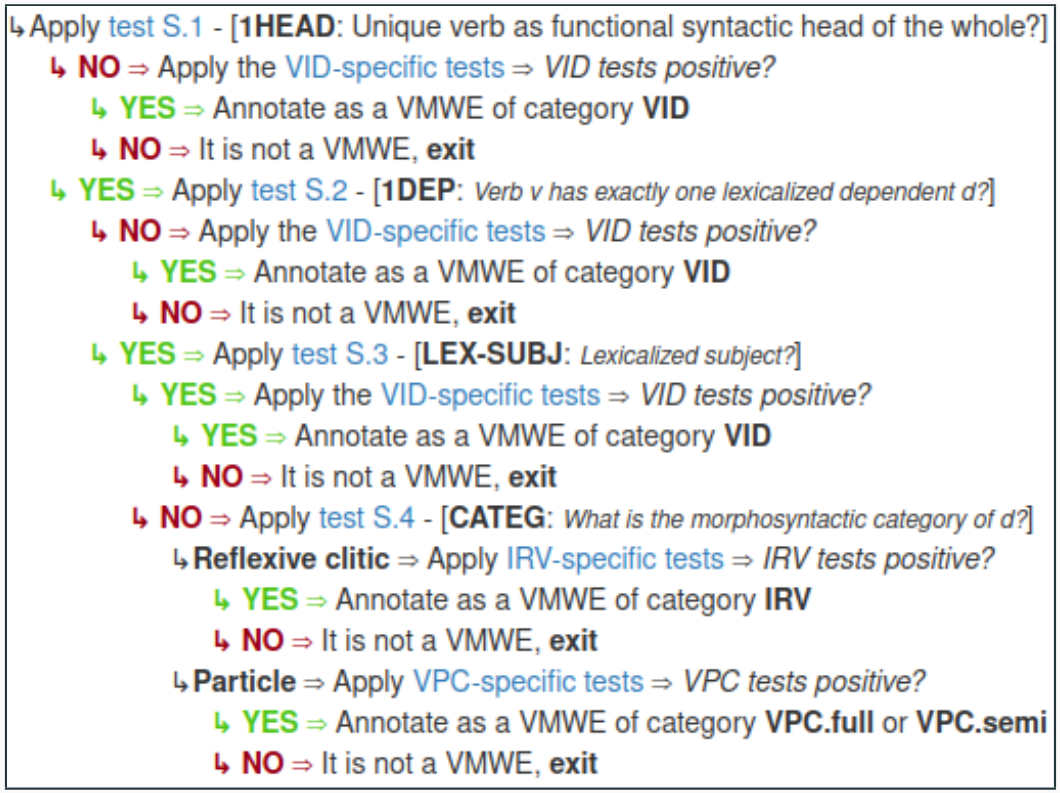
- A document describing the task in detail
  - → Precise definitions of terms
  - → Homogeneous/standard notation
  - → Describe what may seem obvious
- Describe corner cases
  - → Borderline or difficult phenomena

# Annotation guidelines example: multiword expressions

Select words belonging to multiword expressions

→ Definitions, span, linguistic tests, priorities, multilingual examples...

→ Decision diagram: chain tests in reproducible annotation "algorithm"




Source: `https://parsemefr.lis-lab.fr/parseme-st-guidelines/`

# How to write (good) guidelines?

- Always keep in mind: who are the annotators?
- Pilot annotation phases
  - → Versioning and changelogs
- As objective as possible
  - → Yes/no tests, decision trees, flowcharts
- Cover as many borderline cases as possible
  - → Arbitrary but consistent decision, discard if needed
- Add many examples!
  - → Explain the expected outcome, step by step

# Step 3: Annotation interface/platform

- Existing platform: install and adapt
  - → Text: Inception, WebAnno, brat, FLAT, Arborator-Grew,...
- DIY: no existing platform OR no need for complex interface
  - → Easy: spreadsheet, txt files, Google form,[2]...
  - → Harder: website (PHP, Java), web framework (Dash, Streamlit),...

| MWE | sentence-with-mweoccur | annotation | comment |
|---|---|---|---|
| abrir vantagem | Após a primeira parcial ficar empatada em 7 a 7 , o Brasil [abriu] uma [vantagem] decisiva com quatro | NOT TO ANNOTATE | NOT TO ANN |
| abster se | Em outro caso , a Quarta Turma manteve decisão que condenou franqueados de a Rede Wizard a [se | NOT TO ANNOTATE | NOT TO ANN |
| acabar se | Isso vale dizer que tendo somente um jogador de razoável condição técnica em o meio , [se] este for r | 5. WRONG-LEXEMES | |
| acabar se | Não importa se você namora há anos , meses ou [se] [acabou] de conhecer o cara . | 5. WRONG-LEXEMES | |
| acabar se | Eles são trabalhadores que lidam com o público e [acabam] [se] tornando confidentes . | 6. COINCIDENTAL | |
| acabar se | Em o Brasil , a iguaria foi trazida por os portugueses e [acabou] [se] popularizando durante a fase Col | 6. COINCIDENTAL | |
| acabar se | Mas o tempo que ele precisará dedicar a sua academia [acabou] [se] tornando um empecilho . | 6. COINCIDENTAL | |
| acabar se | A Iugoslávia [acabou] [se] desintegrando . | 6. COINCIDENTAL | |
| acabar se | Tem gente que a o menor tropeço , desata um rosário de queixas , colocando a culpa em os outros e [ | 6. COINCIDENTAL | |
| acabar se | O príncipe - herdeiro [acabou] casando - [se] com a princesa Margarida de Saboia , sua prima em prim | 6. COINCIDENTAL | |
| acabar se | Vem de lá , em o balanço de o mar / Sob a divina proteção de Iemanjá , odoyá ! / Conduzindo minha e | NOT TO ANNOTATE | NOT TO ANN |
| acabar se | [Acabou] - [se] a Olimpíada , mas a vibração continua fora de os campos e de as raias olímpicas . | NOT TO ANNOTATE | NOT TO ANN |
| acabar se | A tropa está doente e [se] [acabando] . " | NOT TO ANNOTATE | NOT TO ANN |
| acertar a mão | Um subtenente reformado de a Aeronáutica resistiu a a prisão , [acertou] um tiro em [a] [mão] de um a | 6. COINCIDENTAL | Or maybe "na |
| acertar a mão | Celso Roth [acertou] [a] [mão] e o Grêmio faz campanha . | NOT TO ANNOTATE | NOT TO ANN |

[2] Warning: you share your data with Google!

# Step 4: Recruit & train annotators

- Ideal world: your project has an annotation budget
  - → Recruit annotators among students, experts
  - → Organise training sessions
- Most frequent: no or little budget
  - → Ask colleagues, friends
  - → Trade-off: crowdsourcing

# Crowdsourcing

- Compensate for subjectivity = average over many annotators
  - Amazon Mechanical Turk, Crowdflower, ...
- Make the task simpler - accessible for non experts
  - Remuneration per HIT - Human Intelligence Task
- Data quality
  - Qualification pre-task, spammer filtering
- Ethical aspects: unfair remuneration, hard work, ...

# Gamification

- Games with a purpose (GWAP)
  - Fun, visually attractive
  - Competition
- Background: free annotation

  $\rightarrow$ Players = volunteer annotators

# Step 5: Data quality

- Next section

# Outline

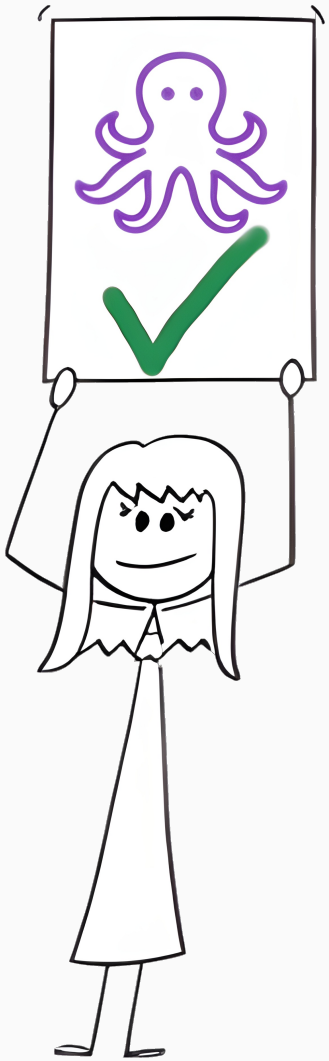
# Double annotation protocol

- Two expert / trained annotators :
  - Same conditions: training, interface, guidelines
  - Annotate the same data items
  - No communication while annotating
- Results should be (almost) identical

  $\longrightarrow$ Do annotators agree at all?

  $\longrightarrow$ How much do they agree/disagree?

# Annotation protocol: example



Manon

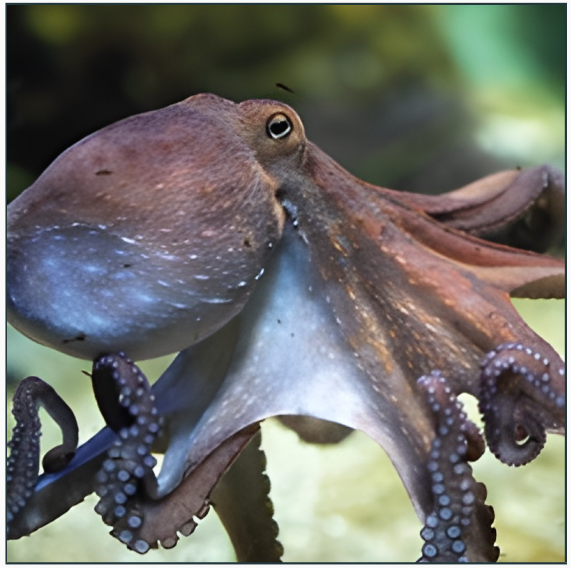
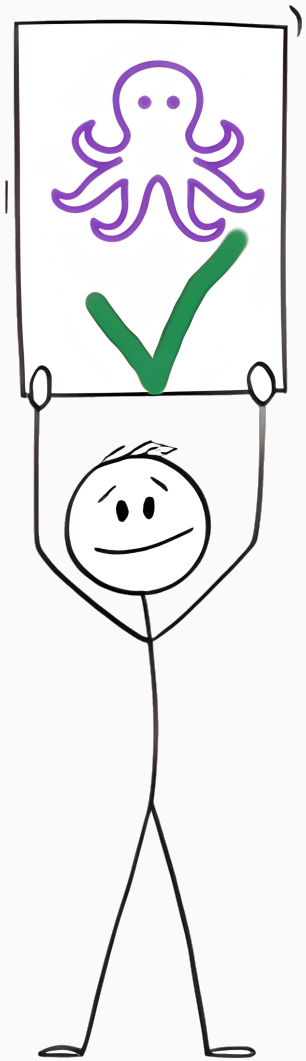
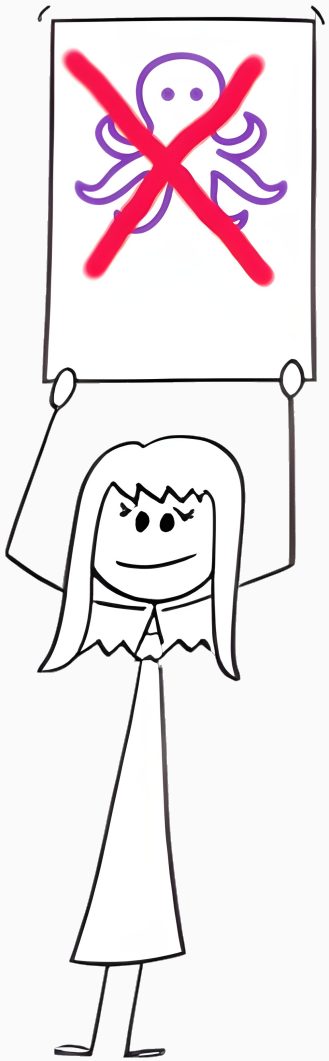


Image $i_1$



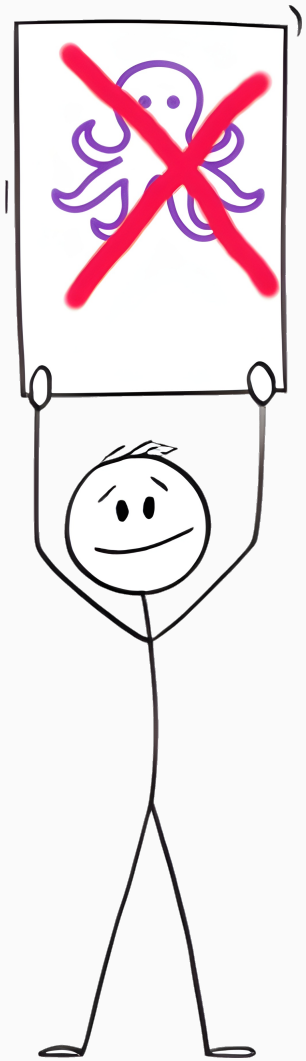
Carlos

# Annotation protocol: example



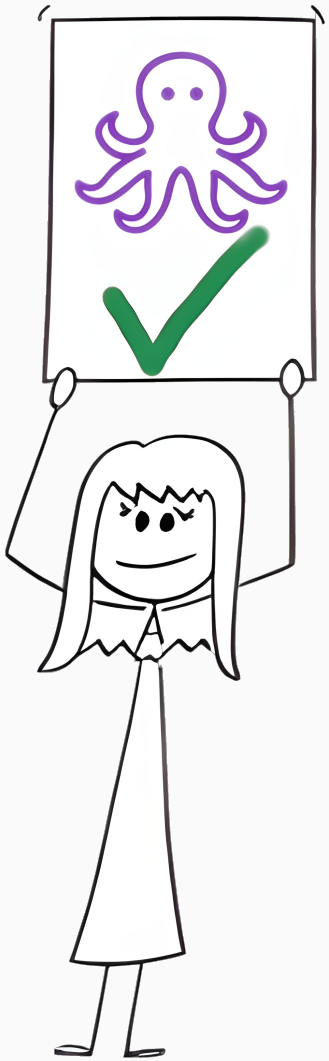
Manon

Image $i_2$



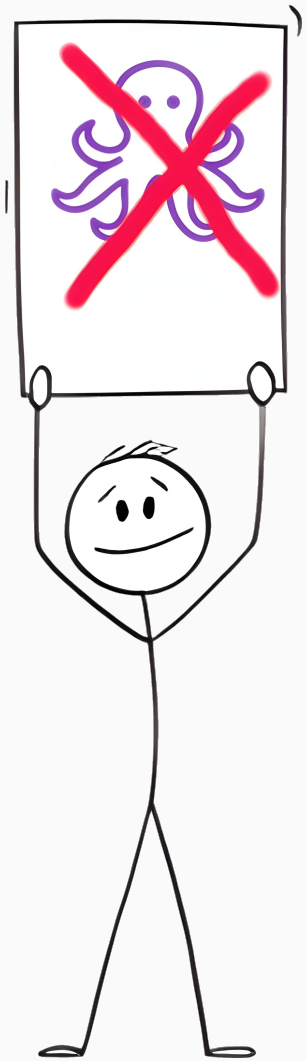
Carlos

# Annotation protocol: example



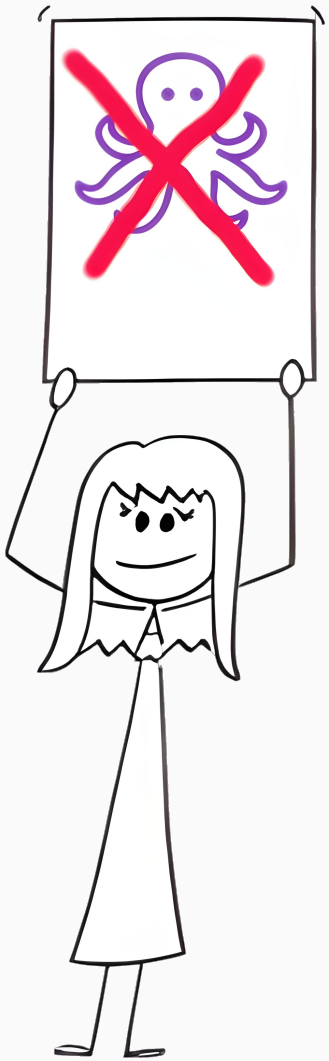
Manon

Image $i_3$



Carlos

# Annotation protocol: example



Manon

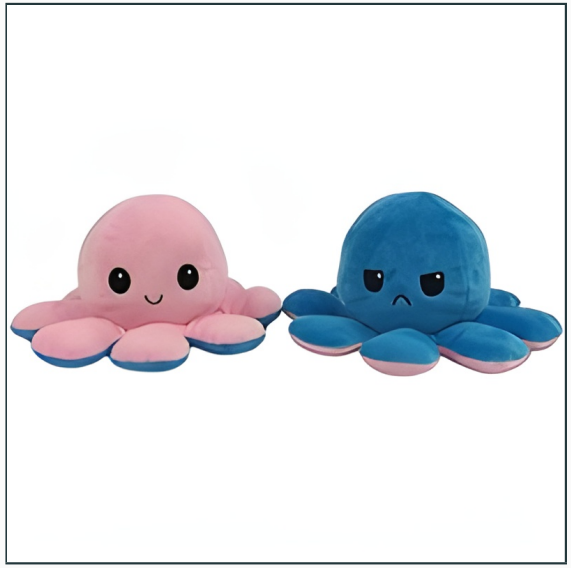
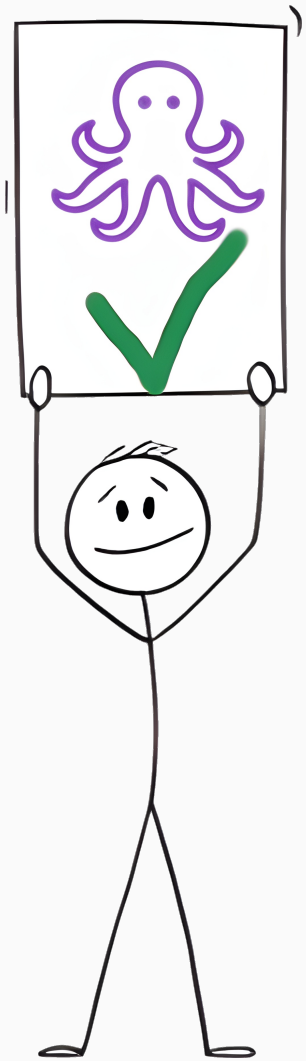


Image $i_4$



Carlos

# Inter-annotator agreement (IAA)

- *High* agreement: I can use my dataset for experiments
- *Low* agreement: restart until agreement is reached

**Inter-annotator agreement**

To what extent can I trust my dataset creation protocol?

→ Are instructions / guidelines clear and objective?

→ Are the annotators well trained?

→ Are the data items of reasonable quality?

# Inter-annotator agreement (IAA)

- *High* agreement: I can use my dataset for experiments
- *Low* agreement: restart until agreement is reached

**Inter-annotator agreement**

To what extent can I trust my dataset creation protocol?

→ Are instructions / guidelines clear and objective?

→ Are the annotators well trained?

→ Are the data items of reasonable quality?

- Quantify "low" and "high" agreement → numerical score

# Inter-annotator agreement: framework

Items, categories and annotators:

- Set of items to annotate: $\{i_k \mid i_k \in I\}$
  - $\rightarrow$ Images of marine animals: $i_1, i_2 \ \ldots \ i_{|I|}$
- Set of categories (labels): $\{c_j \mid c_j \in C\}$
  - $\rightarrow$ $c_1$=octopus, $c_2$=dolphin, $c_3$=shark, ...
- Set of annotators (coders): $\{a_m \mid a_m \in A\}$
  - $\rightarrow$ $a_1$=Manon and $a_2$=Carlos[3]

[3] We will focus on the simplest case of $|A|$=2 annotators for the moment.

# Contingency table / confusion matrix

## Raw annotations

| Image | $a_1$=Manon | $a_2$=Carlos |
|---|---|---|
| $i_1$ | Octopus | Other |
| $i_2$ | Other | Other |
| $i_3$ | Other | Octopus |
| $i_4$ | Octopus | Octopus |
| ... | ... | ... |
| $i_{1000}$ | Other | Other |

$\Longrightarrow$

## Confusion matrix

# Contingency table / confusion matrix

## Raw annotations

| Image | $a_1$=Manon | $a_2$=Carlos |
|---|---|---|
| $i_1$ | Octopus | Other |
| $i_2$ | Other | Other |
| $i_3$ | Other | Octopus |
| $i_4$ | Octopus | Octopus |
| ... | ... | ... |
| $i_{1000}$ | Other | Other |

$\Longrightarrow$

## Confusion matrix

| | Octopus | Other | |
|---|---|---|---|
| Octopus | $n_{11}$ | $n_{12}$ | |
| Other | $n_{21}$ | $n_{22}$ | |
| | | | |

- Categories $c_1$, $c_2$ ... as row/column indices

# Contingency table / confusion matrix

## Raw annotations

| Image | $a_1$=Manon | $a_2$=Carlos |
|---|---|---|
| $i_1$ | Octopus | Other |
| $i_2$ | Other | Other |
| $i_3$ | Other | Octopus |
| $i_4$ | Octopus | Octopus |
| ... | ... | ... |
| $i_{1000}$ | Other | Other |

$\Longrightarrow$

## Confusion matrix

| Manon \ Carlos | Octopus | Other |
|---|---|---|
| Octopus | $n_{11}$ | $n_{12}$ |
| Other | $n_{21}$ | $n_{22}$ |

- Categories $c_1$, $c_2$ ... as row/column indices
- Rows $\rightarrow$ annotator $a_1$; columns $\rightarrow$ annotator $a_2$

# Contingency table / confusion matrix

## Raw annotations

| Image | $a_1$=Manon | $a_2$=Carlos |
|---|---|---|
| $i_1$ | Octopus | Other |
| $i_2$ | Other | Other |
| $i_3$ | Other | Octopus |
| $i_4$ | Octopus | Octopus |
| ... | ... | ... |
| $i_{1000}$ | Other | Other |

$\Longrightarrow$

## Confusion matrix

| Manon \ Carlos | Octopus | Other |
|---|---|---|
| Octopus | 410 | 30 |
| Other | 90 | 470 |

- Categories $c_1$, $c_2$ ... as row/column indices
- Rows $\rightarrow$ annotator $a_1$; columns $\rightarrow$ annotator $a_2$
- Cells $n_{ij}$ count the number of items assigned by $a_1$ to $c_i$ ($i^{th}$ row index) AND by $a_2$ to $c_j$ ($j^{th}$ column index)

# Contingency table / confusion matrix

## Raw annotations

| Image | $a_1$=Manon | $a_2$=Carlos |
|---|---|---|
| $i_1$ | Octopus | Other |
| $i_2$ | Other | Other |
| $i_3$ | Other | Octopus |
| $i_4$ | Octopus | Octopus |
| ... | ... | ... |
| $i_{1000}$ | Other | Other |

$\Longrightarrow$

## Confusion matrix

| Manon \ Carlos | Octopus | Other | Total |
|---|---|---|---|
| Octopus | 410 | 30 | 440 |
| Other | 90 | 470 | 560 |
| Total | 500 | 500 | 1000 |

- Categories $c_1$, $c_2$ ... as row/column indices
- Rows $\rightarrow$ annotator $a_1$; columns $\rightarrow$ annotator $a_2$
- Cells $n_{ij}$ count the number of items assigned by $a_1$ to $c_i$ ($i^{th}$ row index) AND by $a_2$ to $c_j$ ($j^{th}$ column index)
- Last column/row: category distribution for $a_1/a_2$

# Observed agreement $A_O$

- Observed agreement $A_O$: ratio of identically annotated items
- Sum of al $n_{ii}$ cells divided by total number of items $|I|$

  $\rightarrow$ Cells on the diagonal of the confusion matrix

| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

$$A_O =$$

# Observed agreement $A_O$

- Observed agreement $A_O$: ratio of identically annotated items
- Sum of al $n_{ii}$ cells divided by total number of items $|I|$

  $\rightarrow$ Cells on the diagonal of the confusion matrix

| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

$$A_O = \frac{410 + 470}{1000} = 0.88$$

# Wooclap

Wooclap time!

# Correcting for chance

| | | Doctor B | | |
|---|---|---|---|---|
| | | Healthy | Depressed | Total |
| Doctor A | Healthy | 980 | 10 | 990 |
| | Depressed | 10 | 0 | 10 |
| | Total | 990 | 10 | 1000 |

$$A_O =$$

# Correcting for chance

| | | Doctor B | | |
|---|---|---|---|---|
| | | Healthy | Depressed | Total |
| Doctor A | Healthy | 980 | 10 | 990 |
| | Depressed | 10 | 0 | 10 |
| | Total | 990 | 10 | 1000 |

$$A_O = \frac{980 + 0}{1000} = 0.98$$

- Observed agreement biased towards Healthy category
  - No agreement in Depressed category
- Most patients are not depressed

  $\rightarrow$ Both guess "100% Healthy" $\rightarrow$ agree most of the time

  $\rightarrow$ High expected agreement $A_E$

# Observed agreement $A_O$ vs. expected agreement $A_E$

- Observed agreement $A_O$ accounts for two cases:
  1. Annotators are confident and agree
  2. Annotators are unsure and guess - but agree by pure chance!
- Expected agreement $A_E$ describes case 2

  $\rightarrow$ Probability of agreeing by pure chance

# Observed agreement $A_O$ vs. expected agreement $A_E$

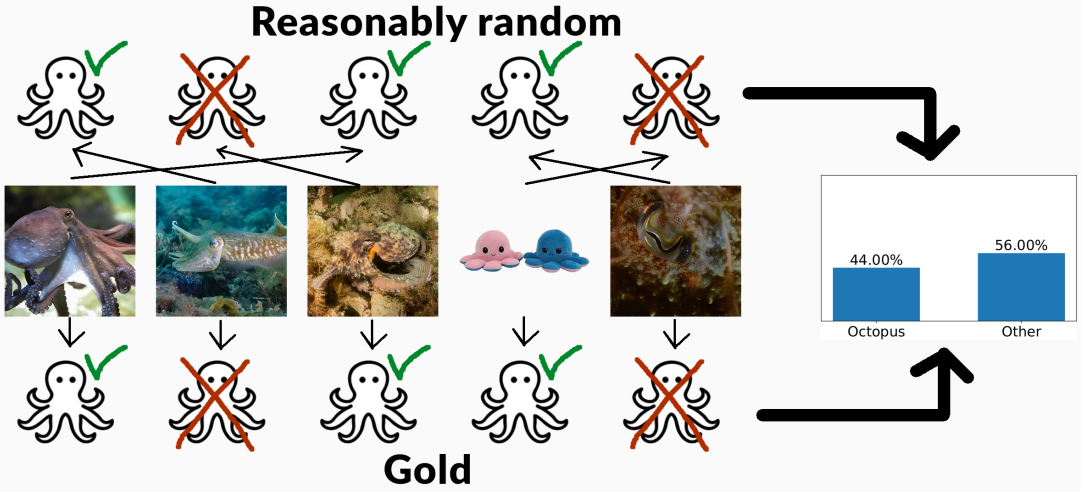
- Observed agreement $A_O$ accounts for two cases:
  1. Annotators are confident and agree
  2. Annotators are unsure and guess - but agree by pure chance!
- Expected agreement $A_E$ describes case 2

  $\rightarrow$ Probability of agreeing by pure chance
- We must correct $A_O$ by discounting $A_E$ from it

  $\rightarrow$ The resulting **chance-corrected score** corresponds to case 1
- How to estimate probability $A_E$ of agreeing by random guess ?

# “Reasonably random” guesses

- Assume annotators never guess 100% at random
- Instead, they respect the category distribution:
  - → The assignment of items to categories is random
  - → But the proportion of each category is reasonable
- Intuition: category distribution influences expected agreement
  - → More probable category ⟹ reasonably random guess agrees more



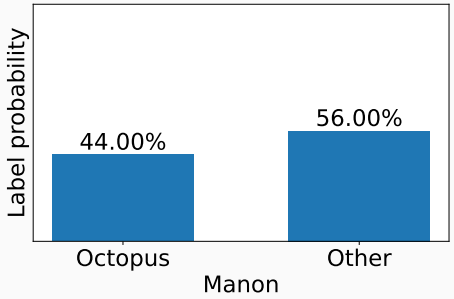
# Wooclap

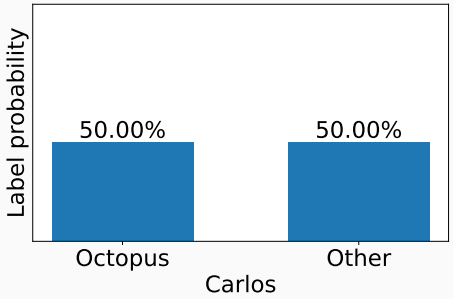
Wooclap time!

# High/low expected agreement

In which situation is $A_E$ higher?

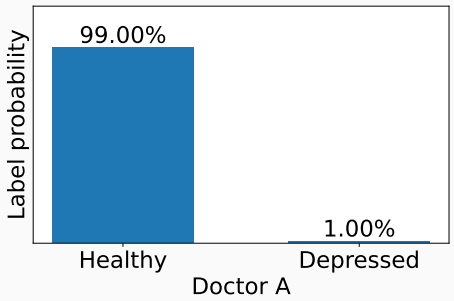
| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

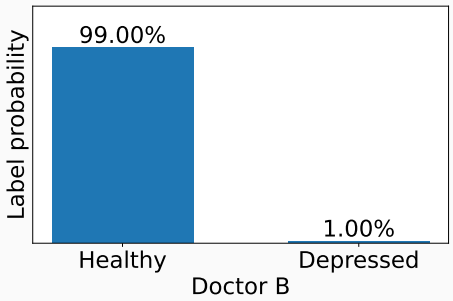



OR

| | | Doctor B | | |
|---|---|---|---|---|
| | | Healthy | Depres. | Total |
| Doctor A | Healthy | 980 | 10 | 990 |
| | Depres. | 10 | 0 | 10 |
| | Total | 990 | 10 | 1000 |

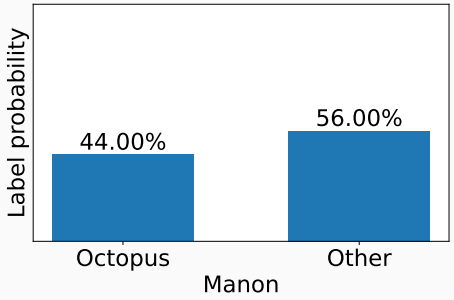


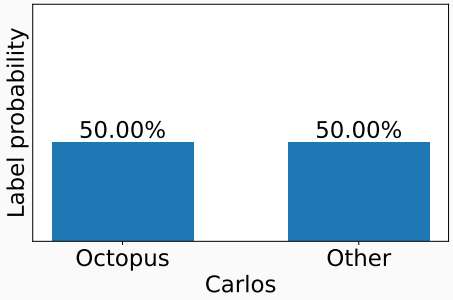



# High/low expected agreement

In which situation is $A_E$ higher?

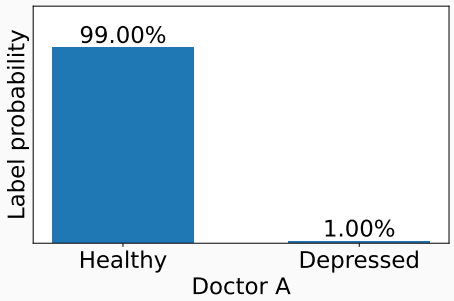
| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

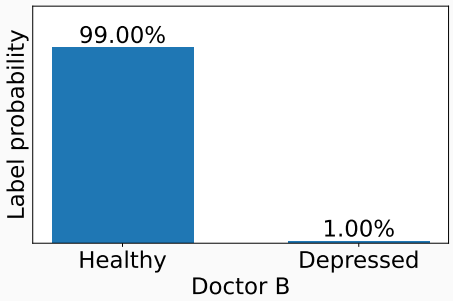



OR

| | | Doctor B | | |
|---|---|---|---|---|
| | | Healthy | Depres. | Total |
| Doctor A | Healthy | 980 | 10 | 990 |
| | Depres. | 10 | 0 | 10 |
| | Total | 990 | 10 | 1000 |




$\rightarrow$ The second situation, involving the doctors !

# Category distribution

- Idea: use the category distribution to calculate $A_E$
- But we don't have 1 overall category distribution, but 2!

  $\rightarrow$ One distribution per annotator
- How to estimate the category distribution from annotations?

# Category distribution

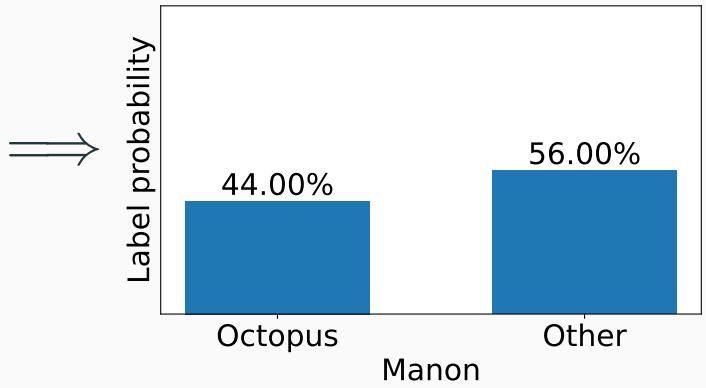
- Idea: use the category distribution to calculate $A_E$
- But we don't have 1 overall category distribution, but 2!
  $\rightarrow$ One distribution per annotator
- How to estimate the category distribution from annotations?
  $\rightarrow$ Combine both annotator's category distributions
  $\rightarrow$ Infer $A_E$ from last row/column of the contingency table

# Category distribution: example

| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

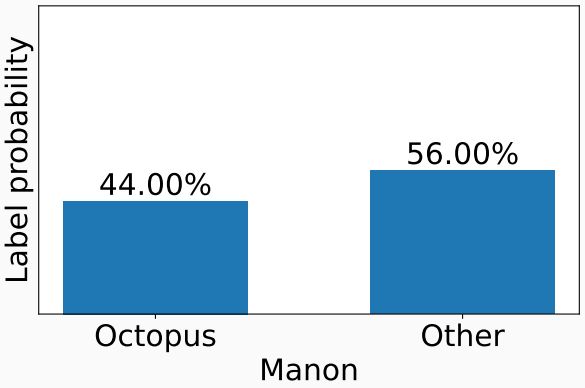
# Category distribution: example

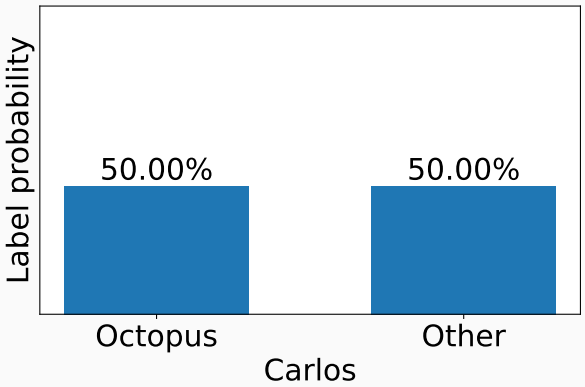
| Manon \ Carlos | Octopus | Other | Total |
|---|---|---|---|
| Octopus | 410 | 30 | 440 |
| Other | 90 | 470 | 560 |
| Total | 500 | 500 | 1000 |



# Category distribution: example

| Manon \ Carlos | Octopus | Other | Total |
|---|---|---|---|
| Octopus | 410 | 30 | 440 |
| Other | 90 | 470 | 560 |
| Total | 500 | 500 | 1000 |



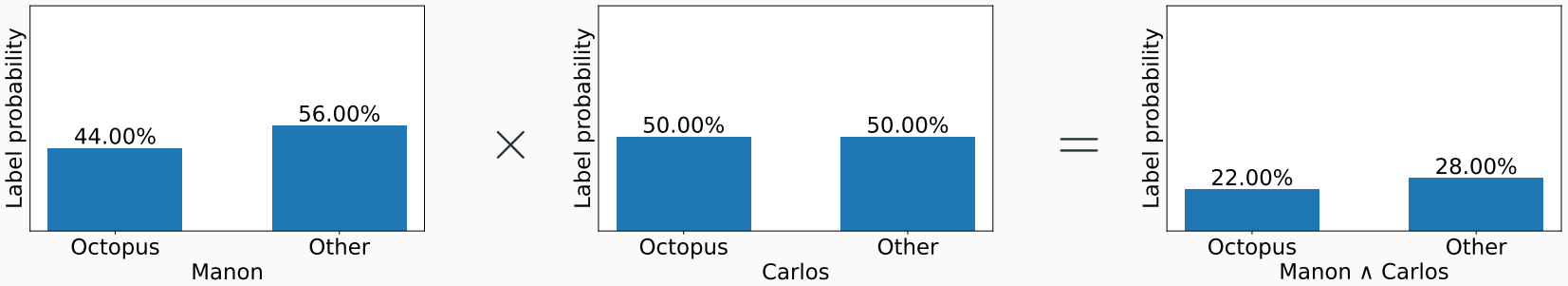

Note: the 50-50 distribution is a coincidence, Carlos did not annotate at random.

# Estimating expected agreement $A_E$

- Independent annotators: joint category distribution = product
  - $\longrightarrow$ Manon and Carlos guess Octopus
  - $\longrightarrow$ Manon and Carlos guess Other

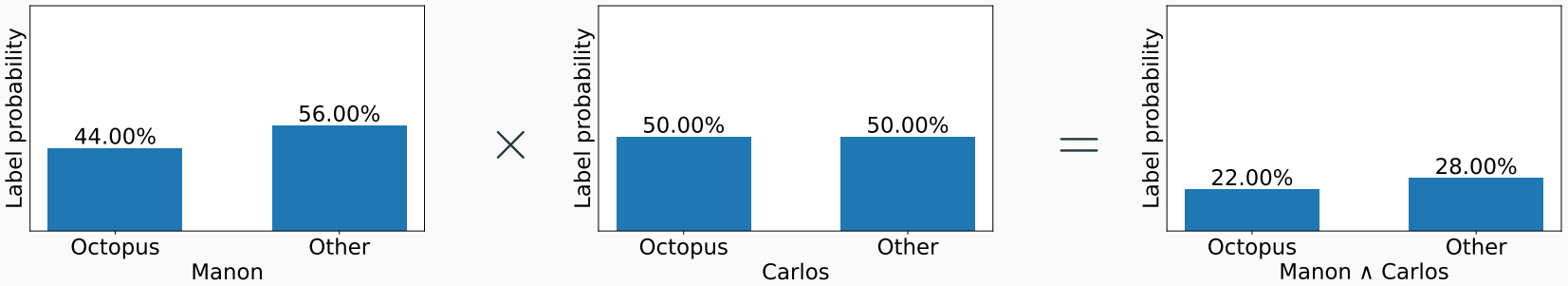
# Estimating expected agreement $A_E$

- Independent annotators: joint category distribution = product
  - $\longrightarrow$ Manon and Carlos guess Octopus
  - $\longrightarrow$ Manon and Carlos guess Other



# Estimating expected agreement $A_E$

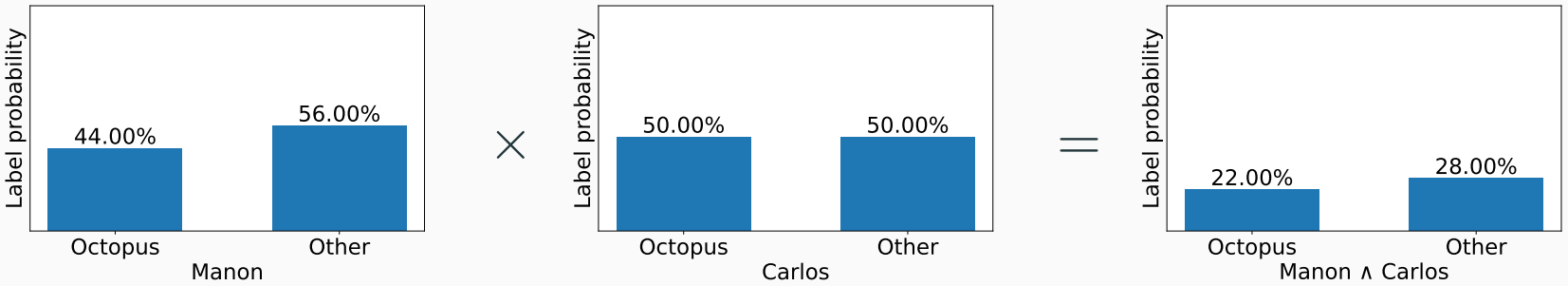
- Independent annotators: joint category distribution = product
  - $\rightarrow$ Manon and Carlos guess Octopus
  - $\rightarrow$ Manon and Carlos guess Other



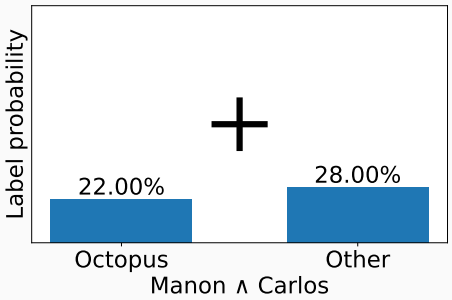

- Mutually exclusive categories: law of total probability= sum
  - $\rightarrow$ Both agree on Octopus or both agree on Other

# Estimating expected agreement $A_E$

- Independent annotators: joint category distribution = product
  - $\rightarrow$ Manon and Carlos guess Octopus
  - $\rightarrow$ Manon and Carlos guess Other




- Mutually exclusive categories: law of total probability= sum
  - $\rightarrow$ Both agree on Octopus or both agree on Other




$$= 0.22 + 0.28 = 0.50 = A_E$$

# Expected agreement $A_E$

1. Get (marginal) category probability distributions per annotator
   - Annotator 1 – normalised $i^{th}$ row sum: $\sum_{j=1}^{|C|} \frac{n_{ij}}{|I|}$
   - Annotator 2 – normalised $i^{th}$ column sum: $\sum_{j=1}^{|C|} \frac{n_{ji}}{|I|}$

# Expected agreement $A_E$

1. Get (marginal) category probability distributions per annotator
   - Annotator 1 – normalised $i^{th}$ row sum: $\sum_{j=1}^{|C|} \frac{n_{ij}}{|I|}$
   - Annotator 2 – normalised $i^{th}$ column sum: $\sum_{j=1}^{|C|} \frac{n_{ji}}{|I|}$
2. Multiply both to estimate overall category distribution

$$\frac{1}{|I|} \sum_{j=1}^{|C|} n_{ij} \times \frac{1}{|I|} \sum_{j=1}^{|C|} n_{ji}$$

# Expected agreement $A_E$

1. Get (marginal) category probability distributions per annotator
   - Annotator 1 – normalised $i^{th}$ row sum: $\sum_{j=1}^{|C|} \frac{n_{ij}}{|I|}$
   - Annotator 2 – normalised $i^{th}$ column sum: $\sum_{j=1}^{|C|} \frac{n_{ji}}{|I|}$
2. Multiply both to estimate overall category distribution

$$\frac{1}{|I|} \sum_{j=1}^{|C|} n_{ij} \times \frac{1}{|I|} \sum_{j=1}^{|C|} n_{ji}$$

3. Sum estimated agreement probabilities over all categories

$$A_E = \sum_{i=1}^{|C|} \left[ \frac{1}{|I|^2} \left( \sum_{j=1}^{|C|} n_{ij} \right) \times \left( \sum_{j=1}^{|C|} n_{ji} \right) \right]$$

# Calculating $A_E$ with the formula

| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

$$A_E = \sum_{i=1}^{|C|} \left[ \frac{1}{|I|^2} \left( \sum_{j=1}^{|C|} n_{ij} \right) \times \left( \sum_{j=1}^{|C|} n_{ji} \right) \right]$$

# Calculating $A_E$ with the formula

| | | Carlos: Octopus | Carlos: Other | Total |
|---|---|---|---|---|
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | $1000 = \lvert I \rvert$ |

$$A_E = \sum_{i=1}^{|C|} \left[ \frac{1}{|I|^2} \left( \sum_{j=1}^{|C|} n_{ij} \right) \times \left( \sum_{j=1}^{|C|} n_{ji} \right) \right]$$

$$= \sum_{i=1}^{2} \left[ \frac{1}{1000^2} \left( \sum_{j=1}^{2} n_{ij} \right) \times \left( \sum_{j=1}^{2} n_{ji} \right) \right]$$

# Calculating $A_E$ with the formula

| | | Carlos: Octopus | Carlos: Other | Total |
|---|---|---|---|---|
| Manon | Octopus | $n_{11}$ =410 | $n_{12}$ =30 | 440 |
| | Other | $n_{21}$ =90 | $n_{22}$ =470 | 560 |
| | Total | 500 | 500 | 1000=\|I\| |

$$A_E = \sum_{i=1}^{2} \left[ \frac{1}{1000^2} \left( \sum_{j=1}^{2} n_{ij} \right) \times \left( \sum_{j=1}^{2} n_{ji} \right) \right]$$

$$= \sum_{i=1}^{2} \frac{(n_{i1} + n_{i2}) \times (n_{1i} + n_{2i})}{1000^2}$$

# Calculating $A_E$ with the formula

| | | Carlos: Octopus | Carlos: Other | Total |
|---|---|---|---|---|
| Manon | Octopus | $n_{11}$ =410 | $n_{12}$ =30 | 440 |
| | Other | $n_{21}$ =90 | $n_{22}$ =470 | 560 |
| | Total | 500 | 500 | 1000=\|I\| |

$$A_E = \sum_{i=1}^{2} \frac{(n_{i1} + n_{i2}) \times (n_{1i} + n_{2i})}{1000^2}$$

$$= \frac{(n_{11} + n_{12}) \times (n_{11} + n_{21})}{1000^2} + \frac{(n_{21} + n_{22}) \times (n_{12} + n_{22})}{1000^2}$$

# Calculating $A_E$ with the formula

| | | Carlos: Octopus | Carlos: Other | Total |
|---|---|---|---|---|
| Manon | Octopus | $n_{11}$ =410 | $n_{12}$ =30 | 440 |
| | Other | $n_{21}$ =90 | $n_{22}$ =470 | 560 |
| | Total | 500 | 500 | 1000=\|I\| |

$$A_E = \frac{(n_{11} + n_{12}) \times (n_{11} + n_{21})}{1000^2} + \frac{(n_{21} + n_{22}) \times (n_{12} + n_{22})}{1000^2}$$
$$= \frac{(440) \times (500)}{1000^2} + \frac{(560) \times (500)}{1000^2}$$

# Calculating $A_E$ with the formula

| | | Carlos: Octopus | Carlos: Other | Total |
|---|---|---|---|---|
| Manon | Octopus | $n_{11}$ =410 | $n_{12}$ =30 | 440 |
| | Other | $n_{21}$ =90 | $n_{22}$ =470 | 560 |
| | Total | 500 | 500 | 1000=\|I\| |

$$
\begin{aligned}
A_E &= \frac{(n_{11}+n_{12})\times(n_{11}+n_{21})}{1000^2} + \frac{(n_{21}+n_{22})\times(n_{12}+n_{22})}{1000^2} \\
&= \frac{(440)\times(500)}{1000^2} + \frac{(560)\times(500)}{1000^2} \\
&= 0.22 + 0.28 \\
&= 0.5
\end{aligned}
$$

# Wooclap

Wooclap time!

# Expected agreement $A_E$

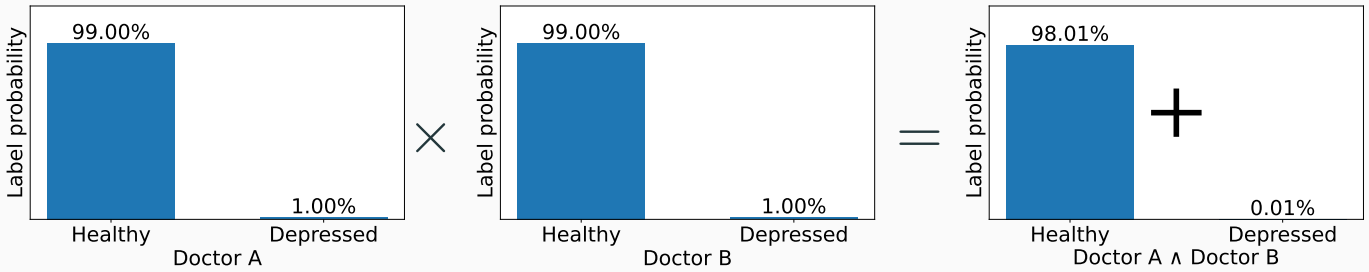
| | | Doctor B | | |
|---|---|---|---|---|
| | | Healthy | Depres. | Total |
| Doctor A | Healthy | 980 | 10 | 990 |
| | Depres. | 10 | 0 | 10 |
| | Total | 990 | 10 | 1000 |

- Calculate the expected agreement $A_E$

# Expected agreement $A_E$

| | | Doctor B | | |
|---|---|---|---|---|
| | | Healthy | Depres. | Total |
| Doctor A | Healthy | 980 | 10 | 990 |
| | Depres. | 10 | 0 | 10 |
| | Total | 990 | 10 | 1000 |

- Calculate the expected agreement $A_E$




$$A_E = \left(\frac{990}{1000} \times \frac{990}{1000}\right) + \left(\frac{10}{1000} \times \frac{10}{1000}\right)$$
$$= (0.99 \times 0.99) + (0.01 \times 0.01)$$
$$= 0.9802$$

# Expected agreement $A_E$

| | | Doctor B | | |
|---|---|---|---|---|
| | | Healthy | Depres. | Total |
| Doctor A | Healthy | 980 | 10 | 990 |
| | Depres. | 10 | 0 | 10 |
| | Total | 990 | 10 | 1000 |

Remember:

$$A_O = \frac{980 + 0}{1000} = 0.98$$

$$A_E = (0.99 \times 0.99) + (0.01 \times 0.01) = 0.9802$$

- Random guessing would lead to roughly same agreement!

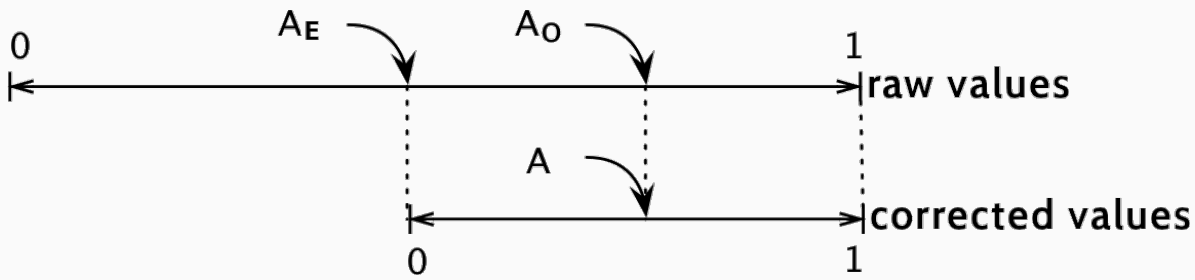
# Kappa: chance-corrected agreement

- Cohen's kappa: proportion of agreement above chance

$$\kappa = \frac{A_O - A_E}{1 - A_E}$$

- Can be seen as a change of scale

$\rightarrow$ e.g. $A_O = .75$ and $A_E = .5 \implies \kappa = .5$



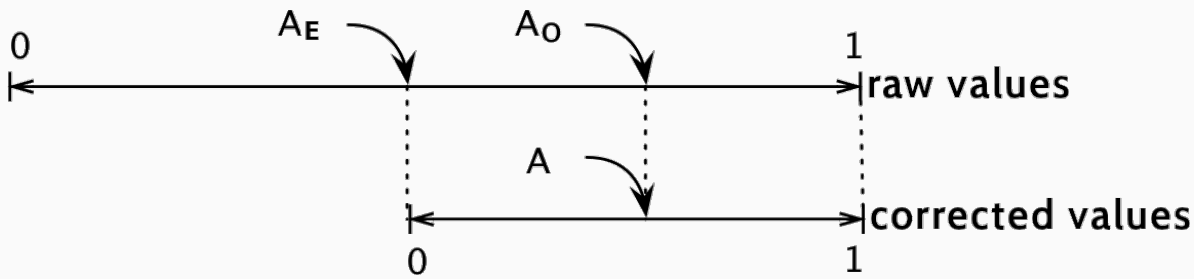
# Kappa: chance-corrected agreement

- Cohen's kappa: proportion of agreement above chance

$$\kappa = \frac{A_O - A_E}{1 - A_E}$$

- Can be seen as a change of scale

  → e.g. $A_O = .75$ and $A_E = .5 \implies \kappa = .5$




- Sanity checks

  → $\kappa > 0$ when $A_O > A_E$

  → $\kappa < 0$ when $A_O < A_E$

Source: Schema adapted from `https://aclanthology.org/2016.tal-2.4/`

# Kappa: example

| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

# Kappa: example

| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

Remember:

$$A_O = 0.88 \qquad A_E = 0.5$$

# Kappa: example

| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

Remember:

$$A_O = 0.88 \qquad A_E = 0.5$$

$$\kappa = \frac{A_O - A_E}{1 - A_E} =$$

# Kappa: example

| | | Carlos | | |
|---|---|---|---|---|
| | | Octopus | Other | Total |
| Manon | Octopus | 410 | 30 | 440 |
| | Other | 90 | 470 | 560 |
| | Total | 500 | 500 | 1000 |

Remember:

$$A_O = 0.88 \qquad A_E = 0.5$$

$$\kappa = \frac{A_O - A_E}{1 - A_E} = \frac{0.88 - 0.5}{1 - 0.5} = 0.76$$

# Wooclap

Wooclap time!

# Calculating kappa: exercise

| | | Doctor B | | |
|---|---|---|---|---|
| | | Healthy | Depres. | Total |
| Doctor A | Healthy | 980 | 10 | 990 |
| | Depres. | 10 | 0 | 10 |
| | Total | 990 | 10 | 1000 |

Remember:

$A_O = 0.98$ $A_E = 0.9802$

1. Calculate Cohen's kappa chance-corrected IAA score

# Calculating kappa: exercise

| | | Doctor B | | |
|---|---|---|---|---|
| | | Healthy | Depres. | Total |
| Doctor A | Healthy | 980 | 10 | 990 |
| | Depres. | 10 | 0 | 10 |
| | Total | 990 | 10 | 1000 |

Remember:

$A_O = 0.98$ $A_E = 0.9802$

1. Calculate Cohen's kappa chance-corrected IAA score

$$\kappa = \frac{0.98 - 0.9802}{1 - 0.9802} = -0.01$$

# Calculating kappa: exercise 2

| | | Annot B | | |
|---|---|---|---|---|
| | | Cat1 | Cat2 | Total |
| Annot A | Cat1 | 0 | 500 | 500 |
| | Cat2 | 500 | 0 | 500 |
| | Total | 500 | 500 | 1000 |

1. Calculate Cohen's kappa chance-corrected IAA score

# Calculating kappa: exercise 2

| | | Annot B | | |
|---|---|---|---|---|
| | | Cat1 | Cat2 | Total |
| Annot A | Cat1 | 0 | 500 | 500 |
| | Cat2 | 500 | 0 | 500 |
| | Total | 500 | 500 | 1000 |

1. Calculate Cohen's kappa chance-corrected IAA score

$$A_O = 0 \quad A_E = 0.5^2 + 0.5^2 = 0.5 \quad \kappa = \frac{0 - 0.5}{1 - 0.5} = -1$$

# Agreement score interpretation

- What is a sufficient kappa value?
  - $\longrightarrow$ 0.67 according to Artstein & Poesio (2008)
  - $\longrightarrow$ 0.75 according to Fleiss (1981)

| | |
|---|---|
| $-1$ to 0 | Less than chance agreement |
| 0 to 0.2 | Slight agreement |
| 0.2 to 0.4 | Fair agreement |
| 0.4 to 0.6 | Moderate agreement |
| 0.6 to 0.8 | Substantial agreement |
| 0.8 to 1 | Almost perfect agreement |

Source: adapted from Landis and Koch (1977)

# Agreement score interpretation

- What is a sufficient kappa value?
  - → 0.67 according to Artstein & Poesio (2008)
  - → 0.75 according to Fleiss (1981)

| | |
|---|---|
| −1 to 0 | Less than chance agreement |
| 0 to 0.2 | Slight agreement |
| 0.2 to 0.4 | Fair agreement |
| 0.4 to 0.6 | Moderate agreement |
| 0.6 to 0.8 | Substantial agreement |
| 0.8 to 1 | Almost perfect agreement |

Source: adapted from Landis and Koch (1977)

- All interpretation scales are **subjective**
- Depends on $|K|$, $|C|$ and $|I| \implies$ significance!
  - → Next chapter...

# Going further

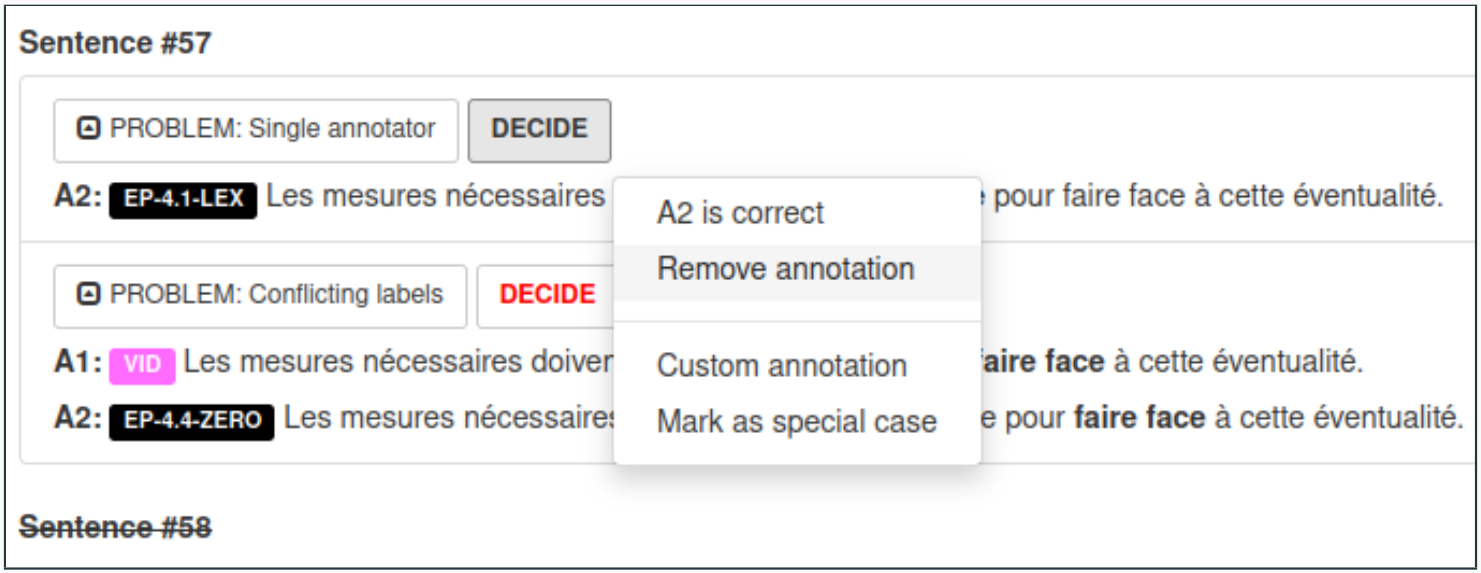
- We only looked at Cohen's kappa, other scores exist ($S$, $\pi$,...)
- More than 2 raters: pairs of agreeing annotations – Fleiss' $\kappa$
- Krippendorf's $\alpha$: weighted by distance between categories
- Sporadic annotations: F-score between raters

**Going further**

- Take a look at `kappa.py` script on the course page

# Step 6: Adjudication / homogenisation

- Creation of (adjudicated) final dataset
- Carried out by another expert (not an annotator)
  - → Dedicated interface
  - → Documented conflict resolution strategies

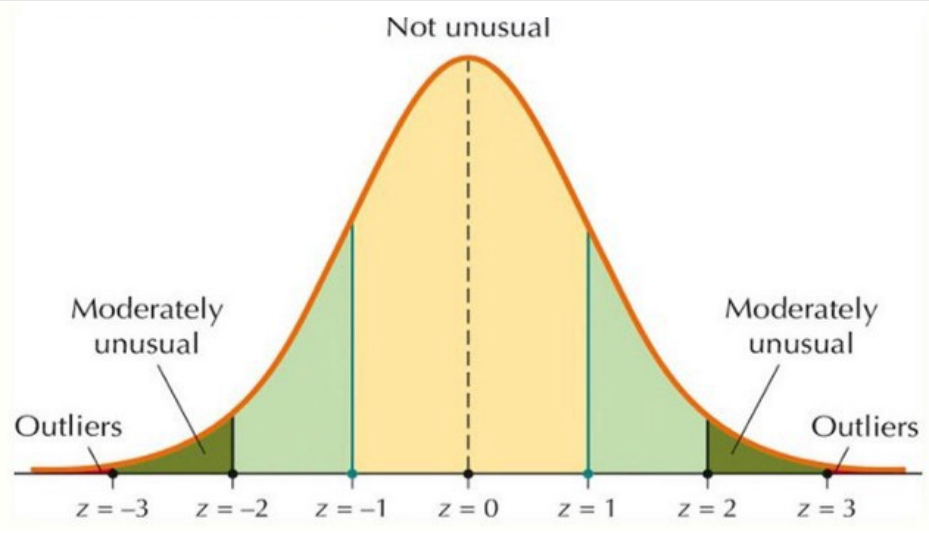


# Data cleaning

- Some annotations are outliers
- Cleaning must occur before experiments

## Z-score filtering

Remove annotations that are more than $z$ standard deviations away from the mean



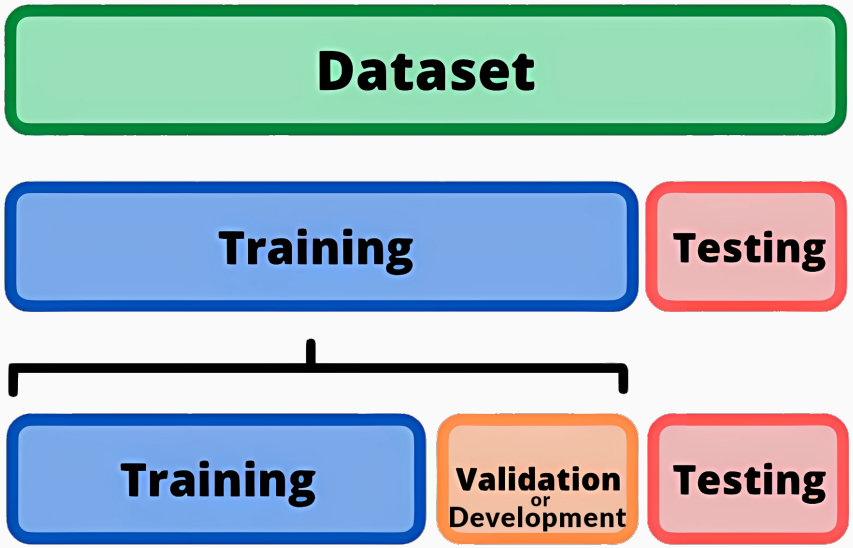



Source: Further reading: https://aclanthology.org/W16-1804/

# Step 7: Data splitting

- Evaluation on held out data: Testing set
- Development on held out data: Development or validation set

  → ⚠ Warning ⚠ it is extremely easy to accidentally tune on test data
- Parameters must be learned from data: Training set



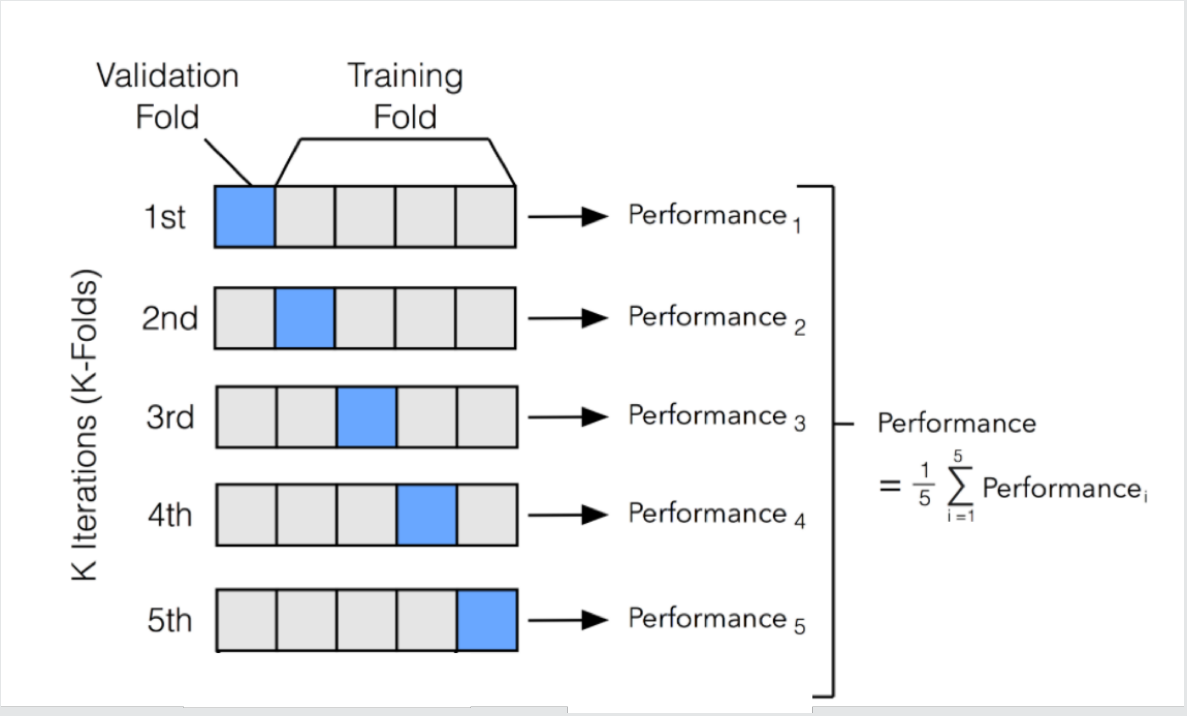
# Splitting strategies i

## Fixed split

- Randomly pick, e.g. 10% for test, 10% for dev, 80% for train
  - $\rightarrow$ Proportions vary according to total dataset size
- Pusblish and use dataset always with the same split
- Comparable across experiments, papers

# Splitting strategies ii

## $k$-fold cross validation




- Expensive: requires training $k$ models instead of 1

# Splitting strategies iii

## Biased split

- Fixed split, but not random
- The test set has controlled characteristics
  - → E.g. test instances are unseen in training data

Discussion:

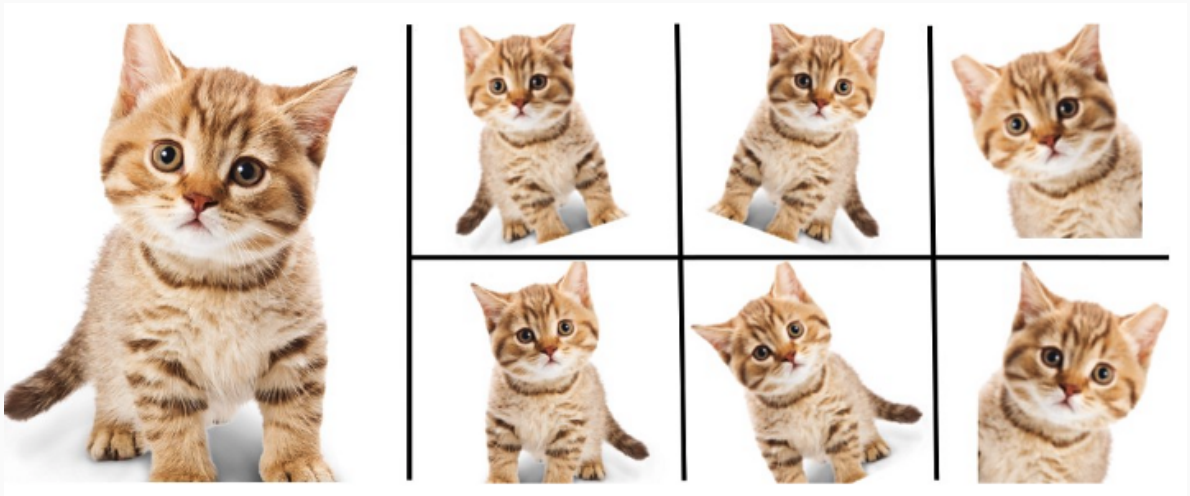
→ Gorman & Bedrick (2019) *We need to talk about standard splits*

→ Søgaard et al. (2021) *We need to talk about random splits*

→ ...

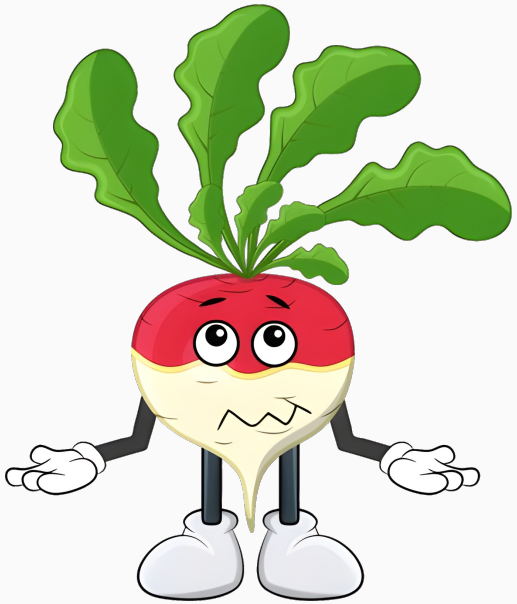
# Dataset size and class imbalance

- Small or imbalanced datasets: hard to learn/evaluate
  - $\longrightarrow$ Large imbalanced: undersampling
  - $\longrightarrow$ Small: augmentation, e.g. rotation, paraphrasing, perturbation...



# Understand the data

- Open your files!
  - → Or take the risk: Y. Goldberg's 2017 post on MILA paper
- Don't try to get blood from a turnip
  - → Maybe your prediction task is unrealistic
  - → Maybe you need other types of resources
  - → ...



# Annotation beyond dataset creation

*"C'est là qu'on voit que la vie n'est pas facile!"*

- Annotating = understanding your problem
  - → Hard for humans? ⟹ maybe hard for models
  - → Low agreement ⟹ maybe ill-defined problem
  - → Annotation guidelines ⟹ inspiration for features

# Outline

# Jumping to conclusions

You develop a neural network B for image classification. You train B on a training set for 50 epochs, evaluate it on a held-out development set, and get an error rate of **8.41%**.

Then, you change the network's architecture by increasing the number of convolution kernels, and train it again for 50 epochs on the same training set. The new model A obtains an error rate of **7.82%** on the development set.

You conclude that A is better than B. However, this conclusion could be wrong.

**What other factors, not taken into account here, can influence the results?**

# Wooclap

Wooclap time!

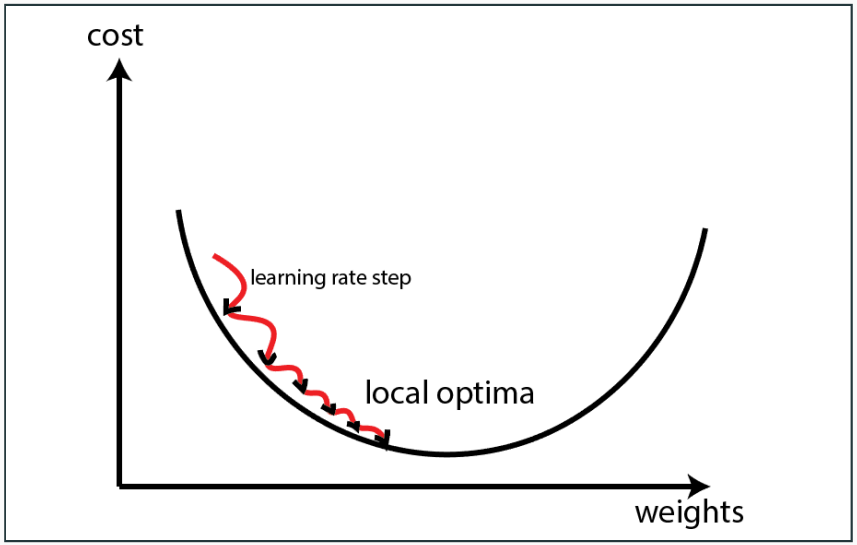
# The devil is in the details

Some details may have great impact on conclusions

1. Hyperparameters
2. Overfitting
3. Model instability
4. Experimental conditions
5. ...

# 1. Hyperparameters (1/2)

- Learning rate
  - $\rightarrow$ Speed at which parameters are updated
- Learning strategy
  - $\rightarrow$ Adam, SGD, warmup steps,...
- Number of epochs
  - $\rightarrow$ Iterations over full dataset



# 1. Hyperparameters (2/2)

- Batch size

  → Fast processing vs. fast convergence

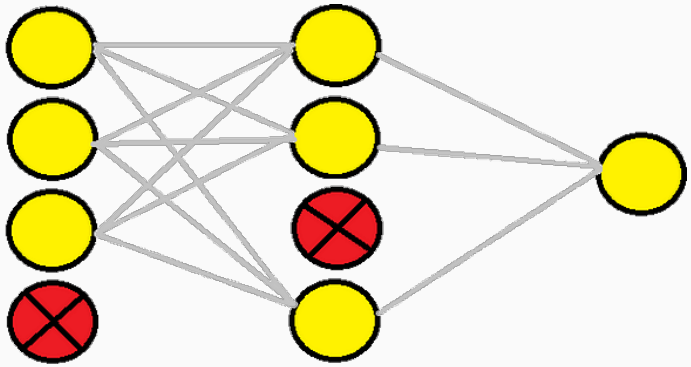
- Model capacity

  → Layer dimensions, nb. of layers, attention heads, conv. filters

- Dropout ratios

  → Prevent memorisation and inefficient parameter use

- ...



# Wooclap

Wooclap time!

# Hyperparameter tuning

- Greedy search (manual search)
  - → Best values found independently from other hyperparameters
- Grid search
  - → All possible (discretised) value combinations
- More sophisticated strategies
  - → Bayesian search, Random search
  - → Specialised libraries: Raytune, optuna, ...
- In practice: intuition and experience help!

# Hyperparameter tuning

- Greedy search (manual search)
  - → Best values found independently from other hyperparameters
- Grid search
  - → All possible (discretised) value combinations
- More sophisticated strategies
  - → Bayesian search, Random search
  - → Specialised libraries: Raytune, optuna, ...
- In practice: intuition and experience help!

**Personal note**

Unavoidable but not very interesting

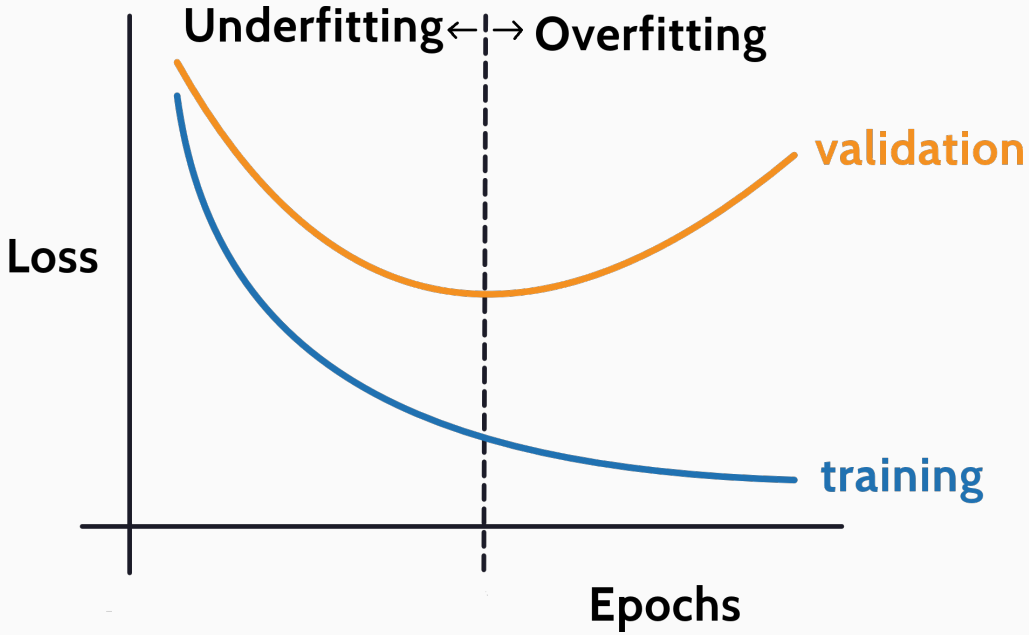
# Wooclap

Wooclap time!

# 2. Overfitting

- The model "overfits" if it memorises the training set
- Rule of thumb of pre-neural models:

  $\rightarrow$ Less features than data items
- Detecting overfit: learning curves on train vs. val set

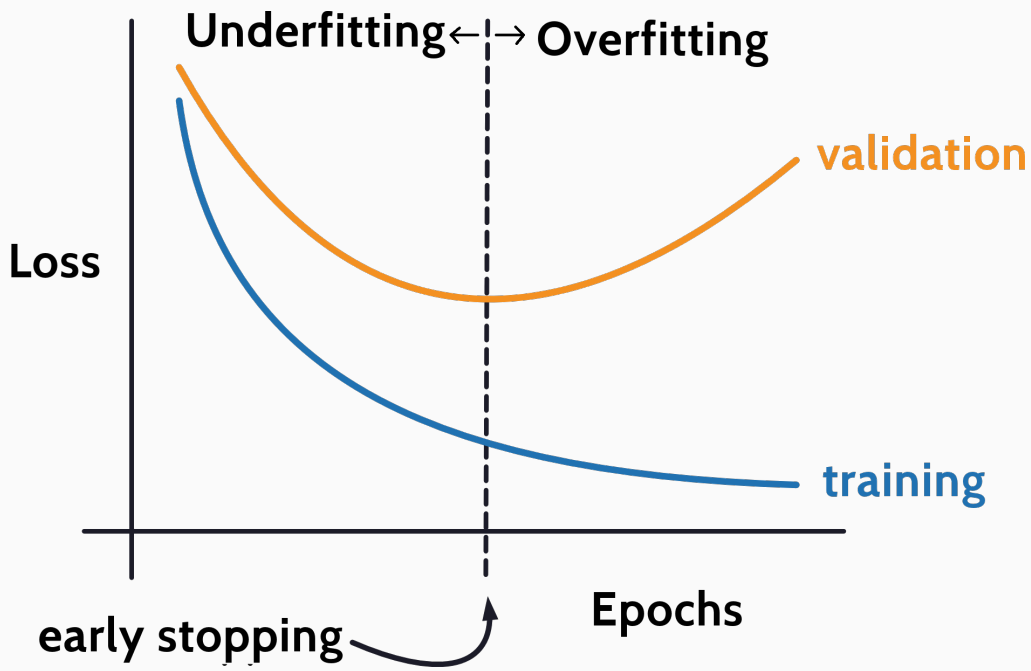


# Early stopping

- After each epoch, calculate loss on val set

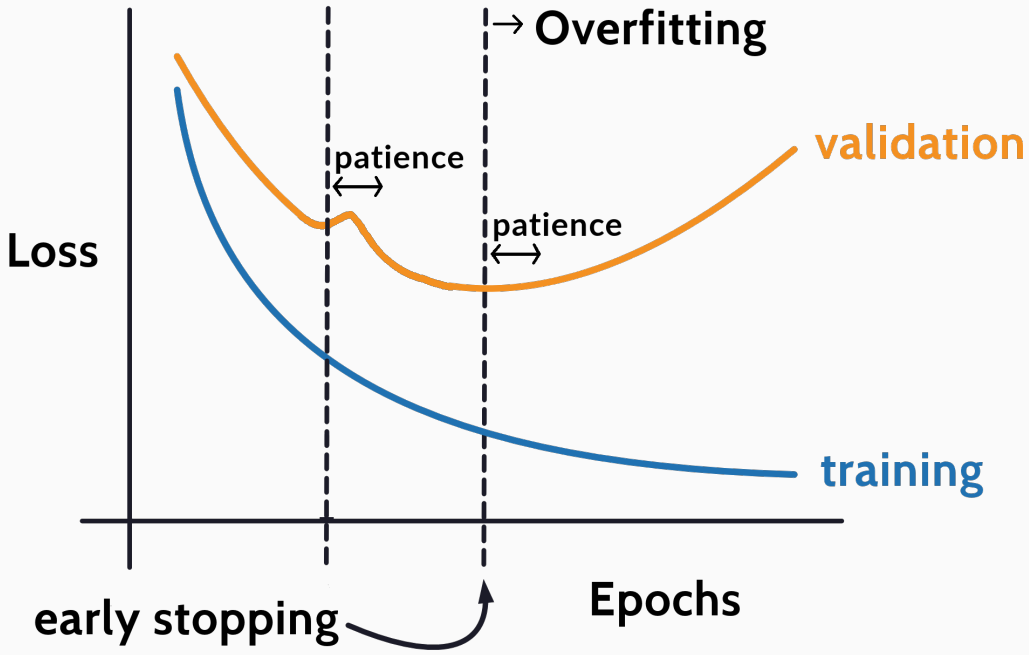
  $\rightarrow$ Loss decreases on training set but not on val set: **stop!**

- Save the model with minimal loss on val set



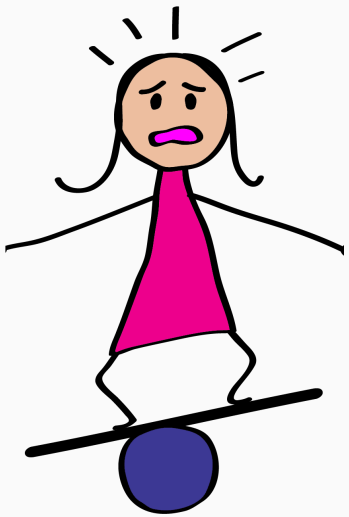
# Patience in early stopping

- Wait a few epochs more (e.g. 3) before deciding to stop
  - → Second chance: maybe val loss will still decrease later



# 3. Model instability

- Same hyperparameters, but different random seeds
  - Parameter initialisation
  - Order of inputs/batches
- Substantially different results
  - Some data orders/initializations consistently better



# Assessing and preventing instability

- Report averages, error bars, confidence intervals
  - Re-run several times with different orders/random seeds
  - Explicitly set `random.seed` (for each lib)
  - Record and publish (in appendix) random seed values
- Early stopping may help
- How to inspect the predictions with several runs?

  $\rightarrow$ Majority vote among predictions

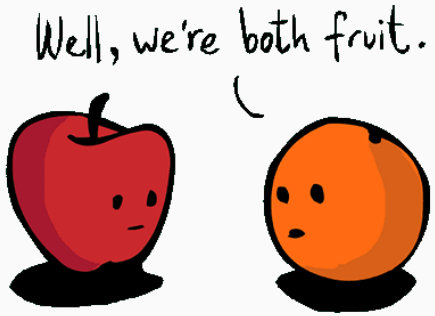
  $\rightarrow$ Ensemble models may improve results!
- Further reading: Dodge et al. (2020)

# 4. Experimental conditions

Experimental conditions can influence conclusions

→ Only compare what is comparable

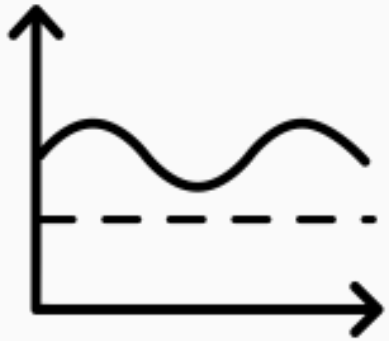
- Amount of supervision
  - → Supervised, unsupervised, semi-supervised
  - → Zero-shot, one-shot, few-shot, ...
- Ablation studies
  - → What part(s) of my model influence which results?



# Baseline

A model is never good or bad per se

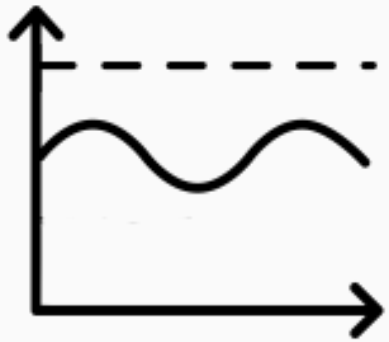
- Situate the model performance wrt. a simpler model
  - → Baseline – simple model for the task
- Examples
  - → Random prediction (too simple?)
  - → Systematically predict most frequent class
  - → A good model 5 years ago
  - → An interpretable model (e.g. decision tree, rules, thresholds)
  - → State-of-the-art model published last month (too complex?)



# Topline

A model is never good or bad per se

- Situate the model performance wrt. a better model
  - → Topline – upper bound for the performance
- Examples
  - → State-of-the-art model published last month
  - → Large model released by big tech company
  - → Human annotator's performance
  - → Same experiment in unrealistic (easy) condition



# It gets messy very fast!

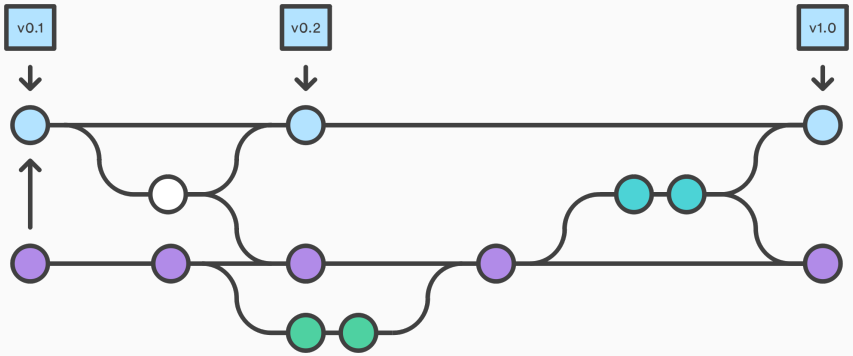
- Logbook
  - Experimental conditions for each result
  - Raw results and links to results
  - Write down ideas, hypotheses, etc.
- Experiments management platform
  - Tensorboard, RayTune, MLFlow, Lightning...

# Distributed systems

- Parallelisation: run experiments on clusters
  - → Distributed supercomputer to run many experiments simultaneously
  - → National: Jean Zay, AMU: mesocentre, local...
- Use of job schedulers: `oar`, `slurm`...
- Organisation is crucial: folder names, log files, timestamps...

# Tools to collaborate

- Git: branches, merge requests, CI for testing
- Overleaf: collaborative LaTeX writing
- Online shared documents: spreadsheets, text documents



# Reproducibility vs. replicability

- Results are reproducible
  - Data available under open licences
  - Model/code shared under open licences
  - Parameters and hyperparameters described
  - Computational requirements reasonable
- Results are replicable
  - Robust to other datasets
  - Robust to different experimental conditions
  - Robust across conditions

Source: ACL 2022 Reproducibility tutorial

# Wooclap

Wooclap time!

# Reprocucibility vs. replicability

A researcher has developed a high-performing model for estimating the mass of a common octopus (*Octopus vulgaris*) from a video taken of it. The method works well for common octopuses shown in various video conditions. One of her colleagues used it to estimate the mass of a blue-ringed octopus (*Hapalochlaena lunulata*). The results are not nearly as good for this species of octopus, in fact they were quite poor.

The initial experiment is:

- Reproducible
- Replicable

# Reprocucibility vs. replicability

A researcher has developed a high-performing model for estimating the mass of a common octopus (*Octopus vulgaris*) from a video taken of it. The method works well for common octopuses shown in various video conditions. One of her colleagues used it to estimate the mass of a blue-ringed octopus (*Hapalochlaena lunulata*). The results are not nearly as good for this species of octopus, in fact they were quite poor.

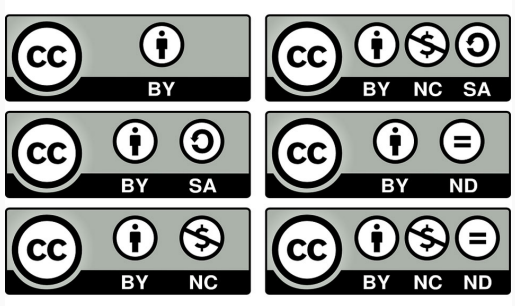
The initial experiment is:

- Reproducible
- Replicable

# Outline

# Licences

- Open science, data sharing, reproducibility
- Data: Creative Commons 4.0
  - SA: share alike
  - NC: non commercial
  - ND: no derivatives
- Code: GNU GPL 2.0
  - Add `LICENCE` file to git repo/zip file
  - Add header to each code file



# Personal data

- Anonymisation
  - Remove all information which allows identifying individuals
  - Aggregate, shuffle
- Pseudo-anonymisation or de-identification
  - Remove identity-related information (name, phone, email)
  - Analysis/crossing could recover individuals identities
- In practice : complete anonymisation barely impossible

# GDPR: general data protection regulation

- Concerns only personal data
- GDPR in a nutshell:
  1. Inform contributors how the data will be used
  2. Provide access and possibility to correct data
  3. Allow data to be removed / forgotten
  4. Inform authorities of any data breach
  5. Ask permission for data use

# GDPR gray zones

- De-identified user-generated content
- Web content retrieved by crawlers/scrapers
- Models pre-trained on personal data
- Web-published artwork (essays, novels, blogs, articles)
    - Copyright vs. GDPR?

# Data repositories

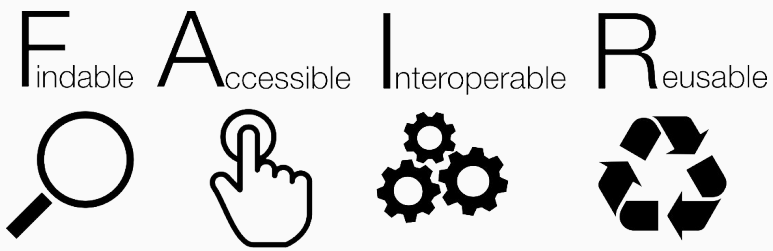
- Temporary: work in progress
  - Public git repo - refer to tags or commit numbers
  - Personal website
  - Consistency can be challenging
  - Backup is important
- Permanent data repository
  - Generic : Zenodo `https://zenodo.org/`

    → Safe, permanent, citable (DOI), free of charge
  - Specialised, e.g. for linguistic datasets:

    → CLARIN-LINDAT, Ortolang, LDC, ELRA, ...

# FAIR principles

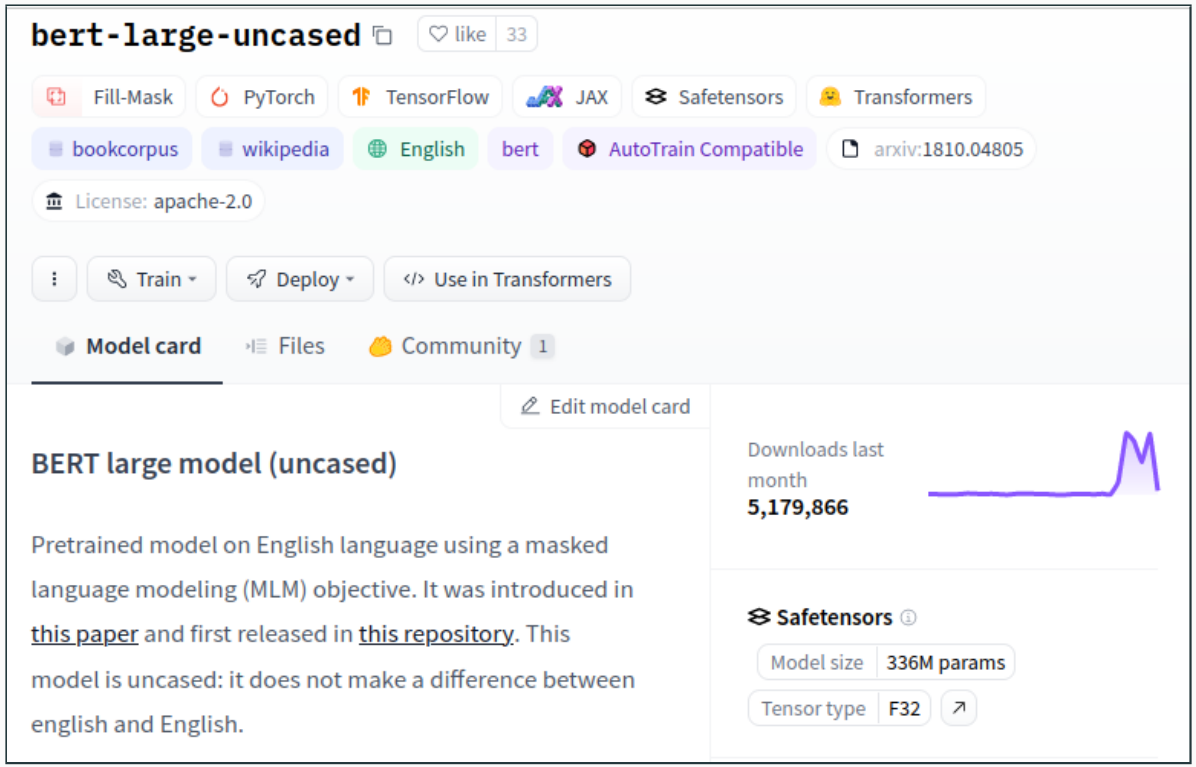
- **F**indable
  - unique ID (DOI, handle.net, URI); present in search engines
- **A**ccessible
  - open protocols/formats for meta-data
- **I**nteroperable
  - well defined, standard, convenient format
- **R**eusable
  - clear licence, document sources, widely adopted standards




Source: FAIR principles – `https://www.go-fair.org/fair-principles/`

# Data/model sheets

- Describe meta-data in standard ways (e.g. Huggingface)




- Further reading: Mitchell et al. (2018) and Gebru et al. (2018)

# Thanks!

## That's all for today

---

Carlos Ramisch
first.last@lis-lab.fr

M2 IAAA - based on the course *Zen Research*
By Carlos Ramisch and Manon Scholivet

# Sources i


- Adeline Paiement's course *Initiation à la recherche*
- Dodge et al. (2020) https://arxiv.org/abs/2002.06305
- Landis & Koch (1977) https://doi.org/10.2307/2529310
- Mathet & Widlöcher (2016) https://aclanthology.org/2016.tal-2.4/
- Mitchell et al. (2018) https://arxiv.org/abs/1810.03993
- Gebru et al. (2018) https://arxiv.org/abs/1803.09010
- Gorman & Bedrick (2019) https://aclanthology.org/P19-1267/
- Poesio & Artstein (2008) https://aclanthology.org/J08-4004/
- Søgaard et al. (2021) https://aclanthology.org/2021.eacl-main.156/
- ACL reproducibility tutorial (2022) https://acl-reproducibility-tutorial.github.io/

# Sources ii

- Creative Commons licence `https://creativecommons.org/`
- FAIR principles `https://www.go-fair.org/fair-principles/`
- GNU GPL licence `https://www.gnu.org/licenses/gpl`
- PARSEME project `https://gitlab.com/parseme/corpora/-/wikis/home`
- Ron Artstein's slides: `http://ron.artstein.org/publications/2012-artstein-agreement-slides.pdf`
- Youtube channels: *DATAtab*, *StatQuest*
- Discussions with Marie Candito, Anna Mosolova, François Hamonic
- Feedback from participants of previous course editions
- Slides illustrated with the help of: Google images, `imgupscaler.com`, Canva

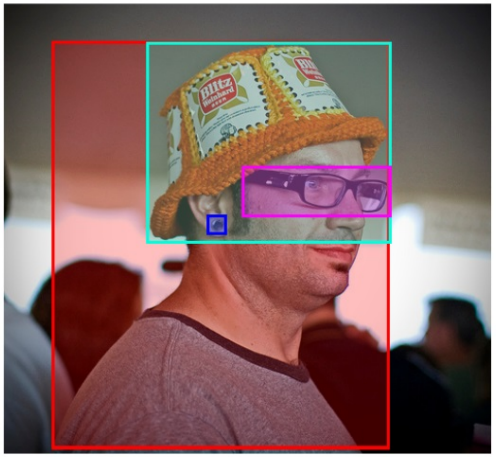
# Sources iii

- Slides written with the help of: ChatGPT, Google Bard, DeepL, Linguee, Overleaf
- Funding: French ANR, through SELEXINI project (ANR-21-CE23-0033-01)

# Backup slides

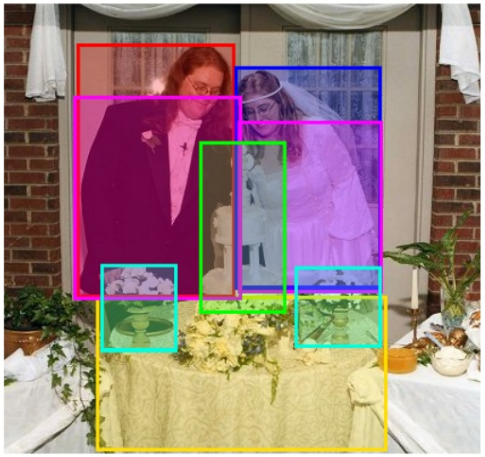
# Indirect annotation example: OpenSubtitles


# Indirect annotation example: Flickr30k (and extensions)



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

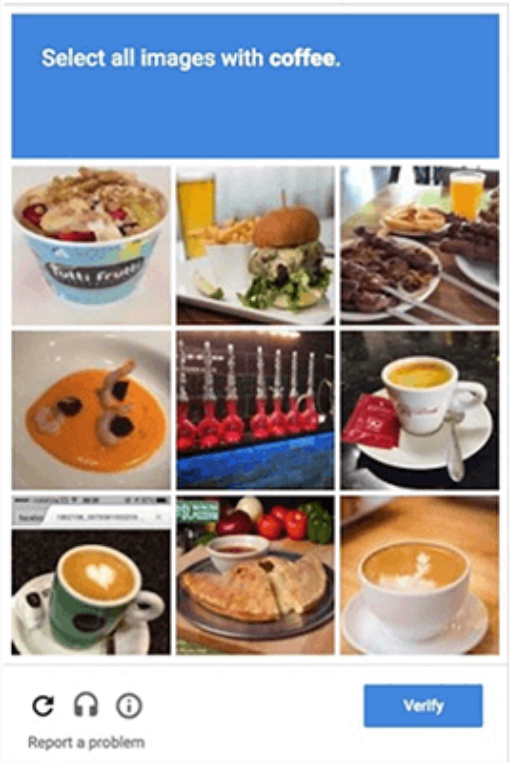
During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .



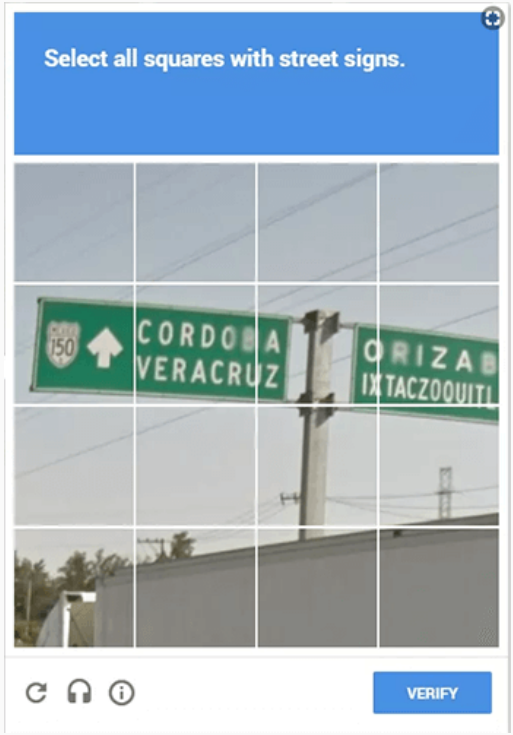
A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

Source: https://bryanplummer.com/Flickr30kEntities/

# Indirect annotation example: Captcha




First one is a captcha...




The second one is free annotation !

# Annotation guidelines example: epidemiology events

## Identify epidemiology events in news

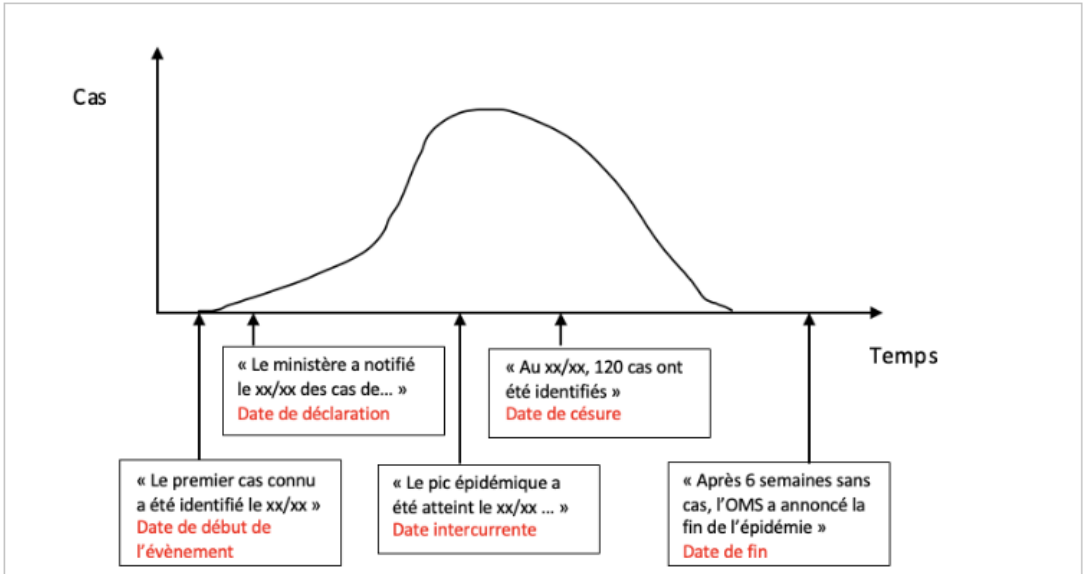
→ date, place, pathology agent, events per document

### 2.2.2. Élément «Date»

Plusieurs types de dates peuvent être retenues :

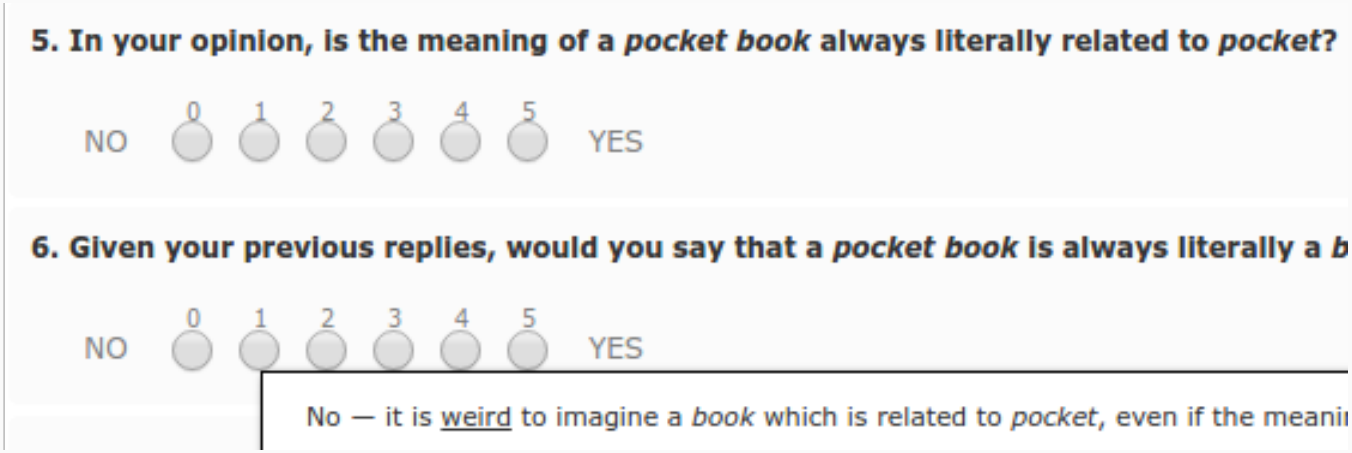
- date de déclaration de l'événement (exemple : "le gouvernement malien a notifié le 13 mai 2020") ;
- date de début de l'événement (exemple : "depuis le début de l'épidémie, le 12 octobre, 123 cas...);
- date de fin de l'événement (exemple : "après 6 semaines sans cas, l'OMS a déclaré le 19 mai 2020 la fin de l'épidémie... ");
- date de césure des données décrivant l'événement (exemple : "Au 14 septembre 2020, 287 cas de dengue ont été diagnostiqués... ");
- date intercurrente (exemple : "le pic épidémique semble avoir été atteint autour du 15 septembre...").

Exemples de dates à l'occasion d'une épidémie :



# Annotation guidelines example: compositionality

- Given a word combination
  - → ***ivory tower*** → privileged situation
- Proportion of whole's meaning predictable from components?
  - → Comp(*ivory_tower*, *ivory*, *tower*) = 10%
- Scale from 0 (totally idiomatic) to 5 (totally compositional)
  - → Head (*book*), modifier (*pocket*), compound (*pocket book*)



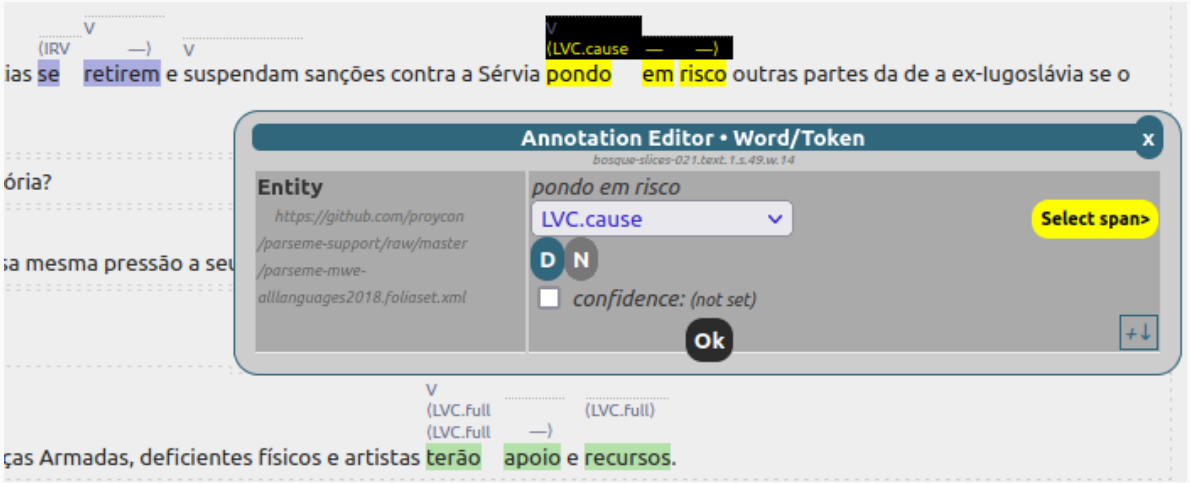
# Annotation interface example: DIY

- Generic tools: Excel spreadseets, text files, etc.
- Web forms from scratch: PHP, Google forms,[4] etc.
- Web dev frameworks: Dash, Streamlit, etc.

| MWE | sentence-with-mweoccur | annotation | comment |
|---|---|---|---|
| abrir vantagem | Após a primeira parcial ficar empatada em 7 a 7 , o Brasil [abriu] uma [vantagem] decisiva com quatro | NOT TO ANNOTATE | NOT TO ANN |
| abster se | Em outro caso , a Quarta Turma manteve decisão que condenou franqueados de a Rede Wizard a [se | NOT TO ANNOTATE | NOT TO ANN |
| acabar se | Isso vale dizer que tendo somente um jogador de razoável condição técnica em o meio , [se] este for r | 5. WRONG-LEXEMES | |
| acabar se | Não importa se você namora há anos , meses ou [se] [acabou] de conhecer o cara . | 5. WRONG-LEXEMES | |
| acabar se | Eles são trabalhadores que lidam com o público e [acabam] [se] tornando confidentes . | 6. COINCIDENTAL | |
| acabar se | Em o Brasil , a iguaria foi trazida por os portugueses e [acabou] [se] popularizando durante a fase Col | 6. COINCIDENTAL | |
| acabar se | Mas o tempo que ele precisará dedicar a sua academia [acabou] [se] tornando um empecilho . | 6. COINCIDENTAL | |
| acabar se | A lugoslávia [acabou] [se] desintegrando . | 6. COINCIDENTAL | |
| acabar se | Tem gente que a o menor tropeço , desata um rosário de queixas , colocando a culpa em os outros e [ | 6. COINCIDENTAL | |
| acabar se | O príncipe - herdeiro [acabou] casando - [se] com a princesa Margarida de Saboia , sua prima em prin | 6. COINCIDENTAL | |
| acabar se | Vem de lá , em o balanço de o mar / Sob a divina proteção de Iemanjá , odoyá ! / Conduzindo minha e | NOT TO ANNOTATE | NOT TO ANN |
| acabar se | [Acabou] - [se] a Olimpíada , mas a vibração continua fora de os campos e de as raias olímpicas . | NOT TO ANNOTATE | NOT TO ANN |
| acabar se | A tropa está doente e [se] [acabando] . " | NOT TO ANNOTATE | NOT TO ANN |
| acertar a mão | Um subtenente reformado de a Aeronáutica resistiu a a prisão , [acertou] um tiro em [a] [mão] de um a | 6. COINCIDENTAL | Or maybe "na |
| acertar a mão | Celso Roth [acertou] [a] [mão] e o Grêmio faz campanha . | NOT TO ANNOTATE | NOT TO ANN |

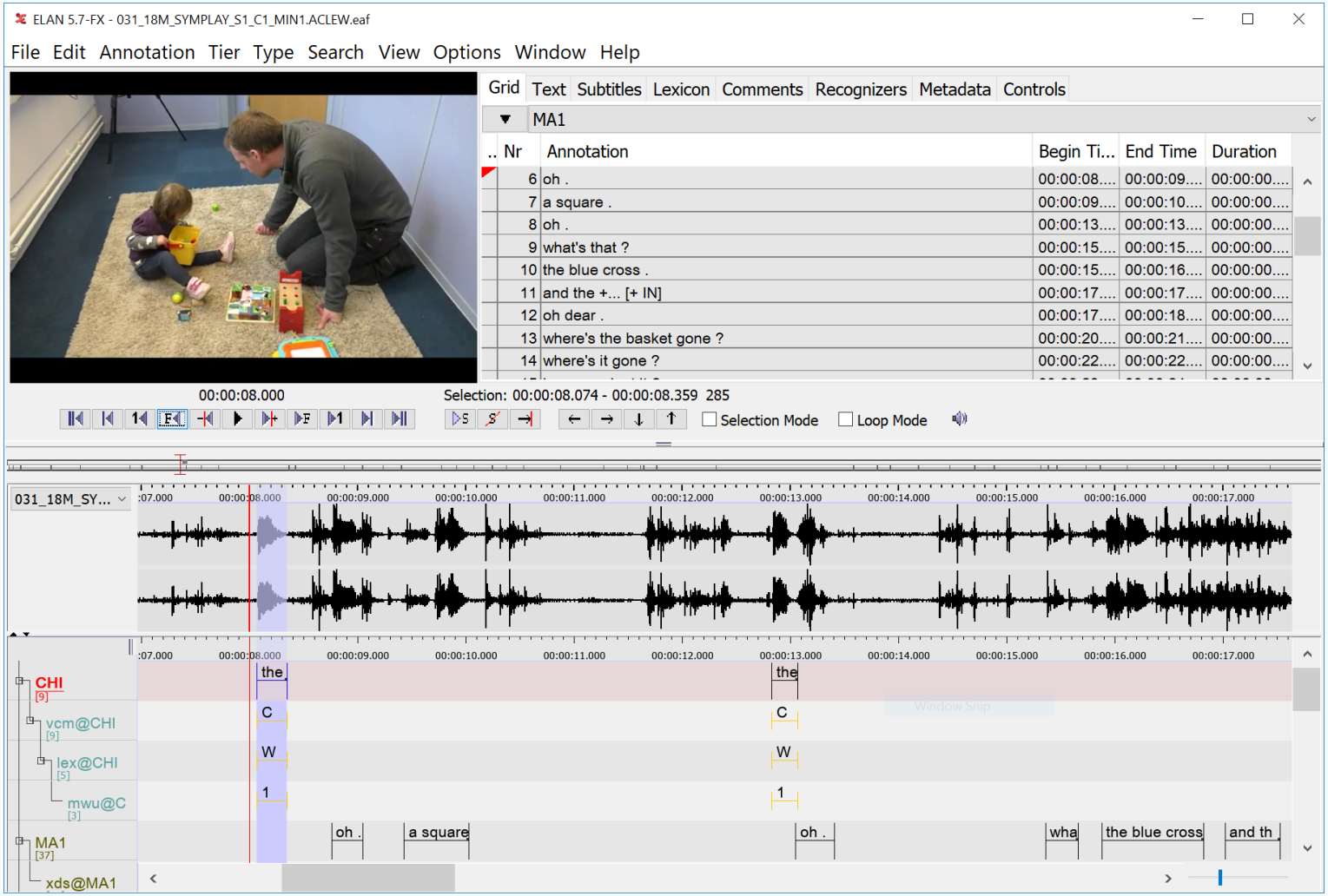
[4] Warning: you share your data with Google!

# Annotation interface example: FLAT (text)



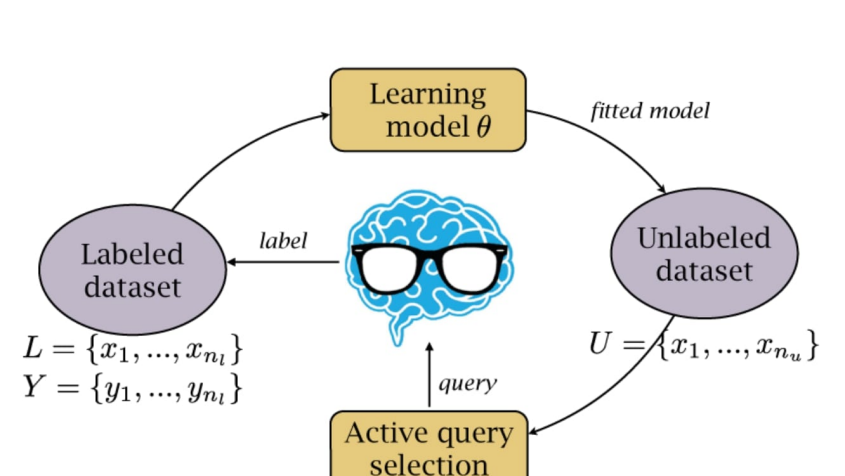

Alternatives: Inception, webAnno, brat, FLAT, Arborator, . . .

# Annotation interface example: ELAN (audio/video)



# Automatic pre-annotation

- Pre-annotation
  1. Annotate a small dataset and train predictive model
  2. Predict on the remaining unlabelled data
  3. Correct the predictions
- **Active learning**
  1. Annotate a given instance
  2. Append to training data and train predictive model
  3. Next instance to annotate chosen automatically
     - Maximise diversity of phenomena
     - Maximise the utility for the model



# Gamification example: Jeux de Mots

Jeux de mots – `https://www.jeuxdemots.org/`

# Gamification example: ZombiLingo

ZombiLingo – `http://gwap.grew.fr/`

# Consistency checks

- Vertical data visualisation
  - Aggregate similar units (e.g. by lemma, POS n-gram, etc)
- Adjudicator of expert annotator corrects mistakes